Data Science and Machine Learning

Prof. Michalis Vlachos

Enterprise for Society

UNIL | Université de Lausanne
HEC Lausanne

IMD
REAL LEARNING. REAL IMPACT

EPFL

The Data Science process

Example 1: What are the demographics of our customers?

Example 2: What is the sentiment about our company in twitter? Mostly positive or negative?

Regression – Predicting numeric values

- Regression is an instance of a **supervised learning** algorithm
- We are given a set of features for some objects/entities (customers, products, etc) AND also the **numeric value** of what we want to predict.

Age	Income	#kids	Withdrawal amount next month
33	100	1	15
55	150	3	21

- Given a new object/entity and its feature values. What would be the numeric value that we want to predict?

38	90	2	??

Applications - Examples

- Given [square meters, number of bathrooms, …] → ? house price

- Given [size of engine, weight of car, …] → ? km/liter of gas

Discussion:

- Find some applications of regression

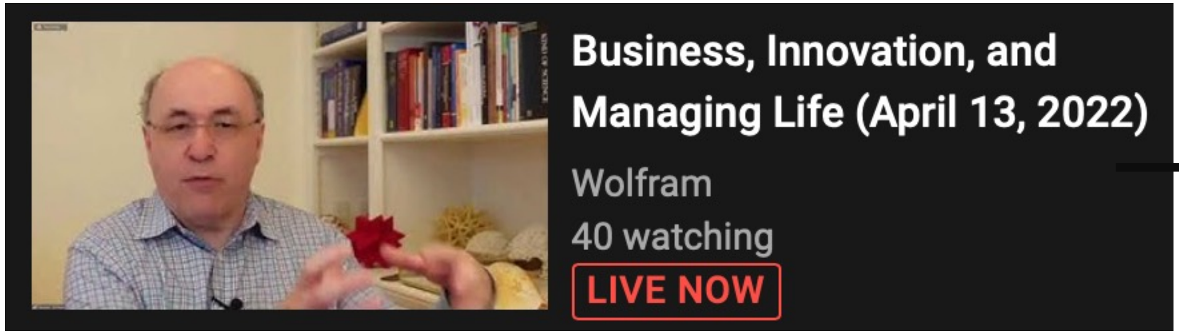
Some examples of regression

- Regression (we predict a number)

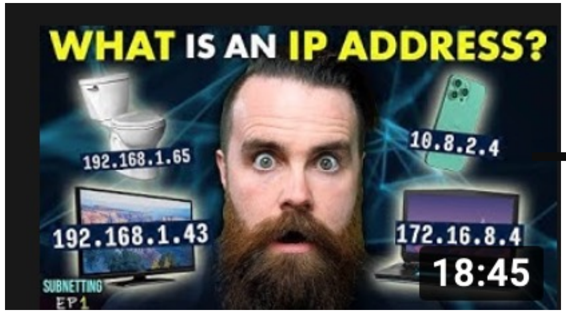
{  → 60 ,  → 42, … } and then when a new example arrived  → ??

Some examples of regression

- Regression (we predict a number)



IQ = 160



IQ = ??

What input features would you use for this application?

Let’s recall (some terminology)

Observation

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75
4	9237-HQITU	Female	0	No	No	2		No	...	151.65

Observation could be a customer, a patient, a car, a country, a novel, a drug, a movie etc

Observation = one row

Feature or attribute

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75
4	9237-HQITU	Female	0	No		2	Yes	No	...	151.65

Feature x = column
(independent variables or predictors)

Feature or attribute

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75
4	9237-HQITU	Female	0	No		2	Yes	No	...	151.65

All the features → X

Feature or attribute

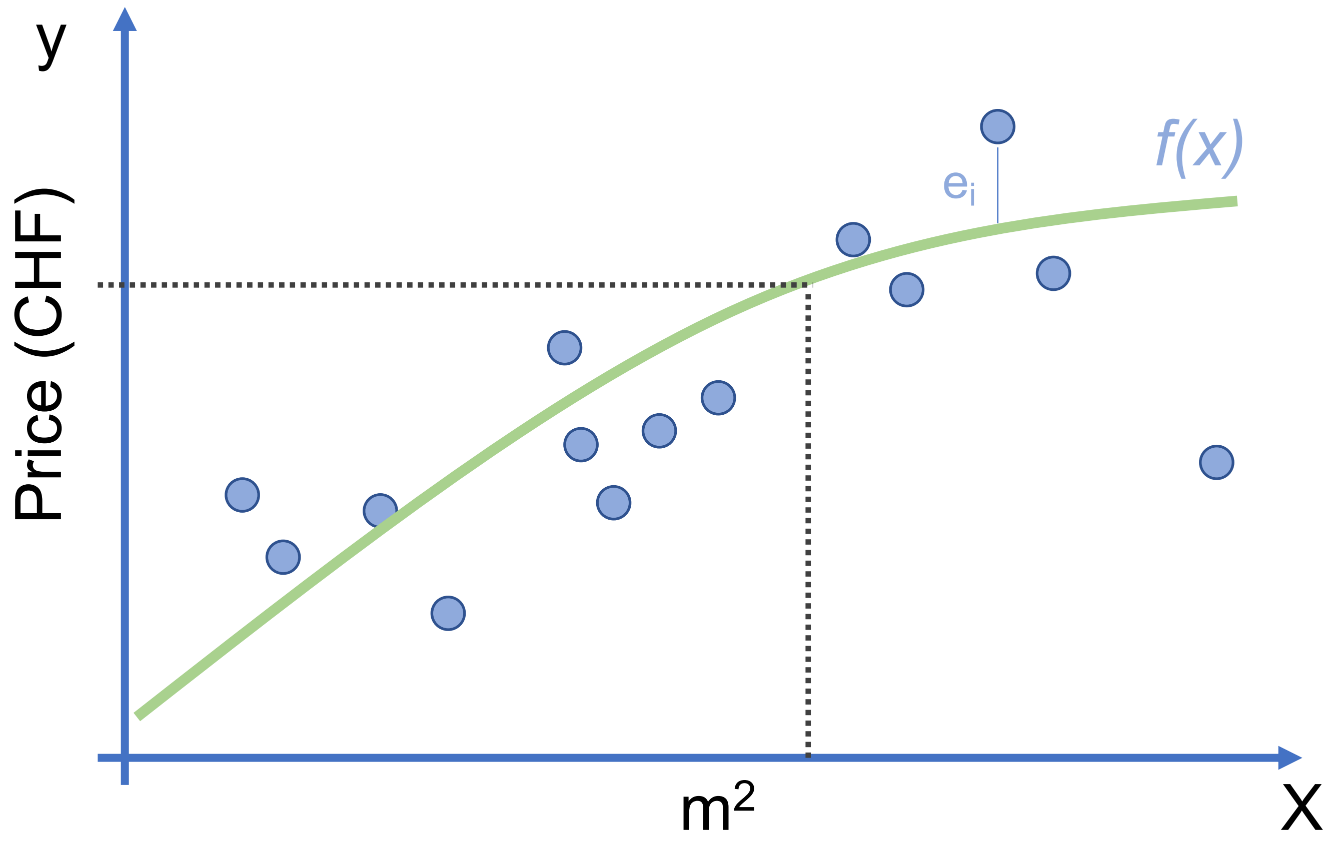
	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	OnlineSecurity	...	TotalCharges
0	7590-VHVEG	Female	0	Yes	No	1	No	No	...	29.85
1	5575-GNVDE	Male	0	No	No	34	Yes	Yes	...	1889.5
2	3668-QPYBK	Male	0	No	No	2	Yes	Yes	...	108.15
3	7795-CFOCW	Male	0	No	No	45	No	Yes	...	1840.75
4	9237-HQITU	Female	0	No	No	2	Yes	No	...	151.65

Target variable y
(dependent variable)

Model representation – Regression

A motivating example

- I would like to sell my house/apartment
- How much should I sell it?

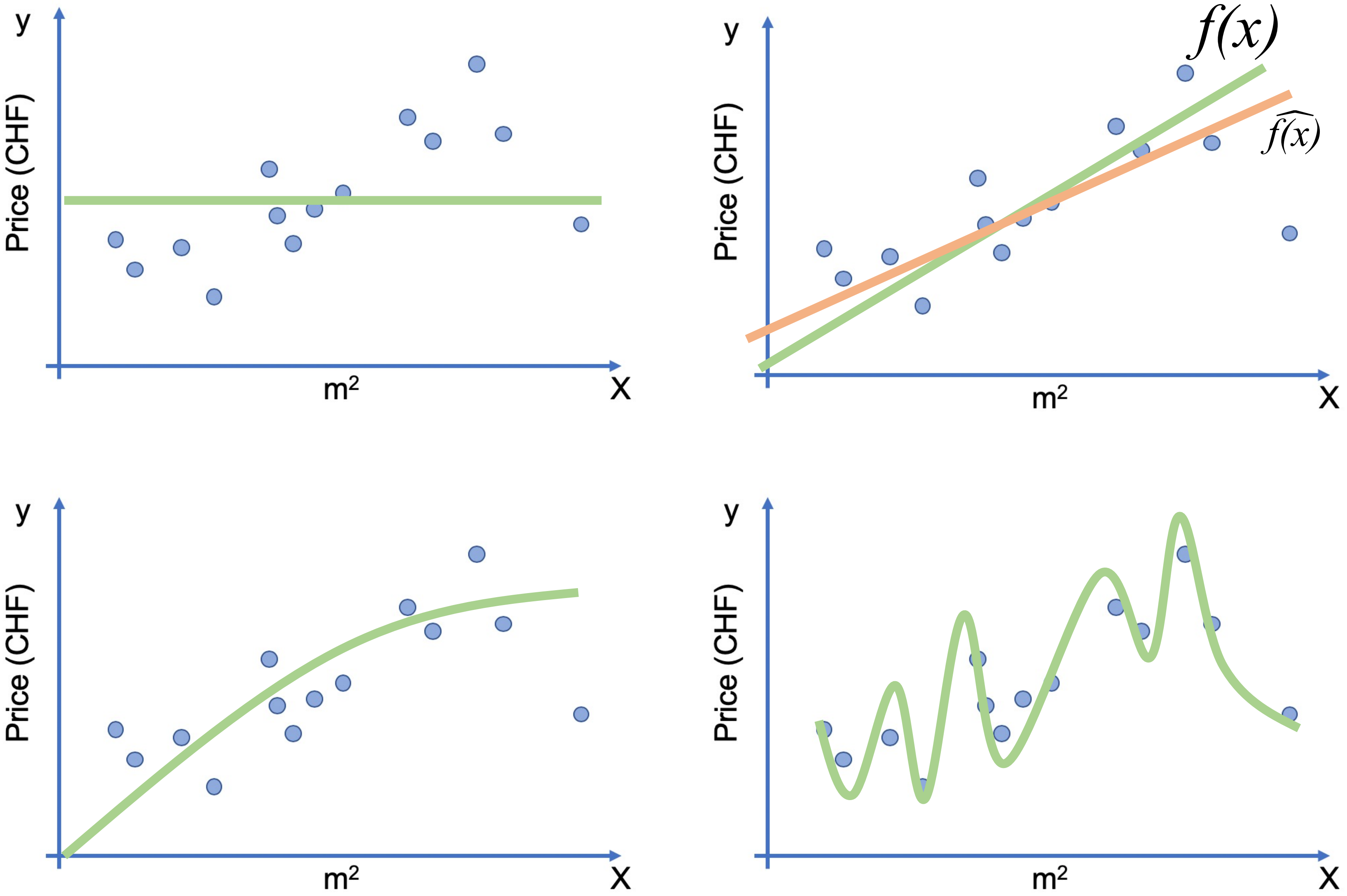


House area	Sale Price
120	930,000
65	705,000
154	2,010,000
…	…

Regression model:

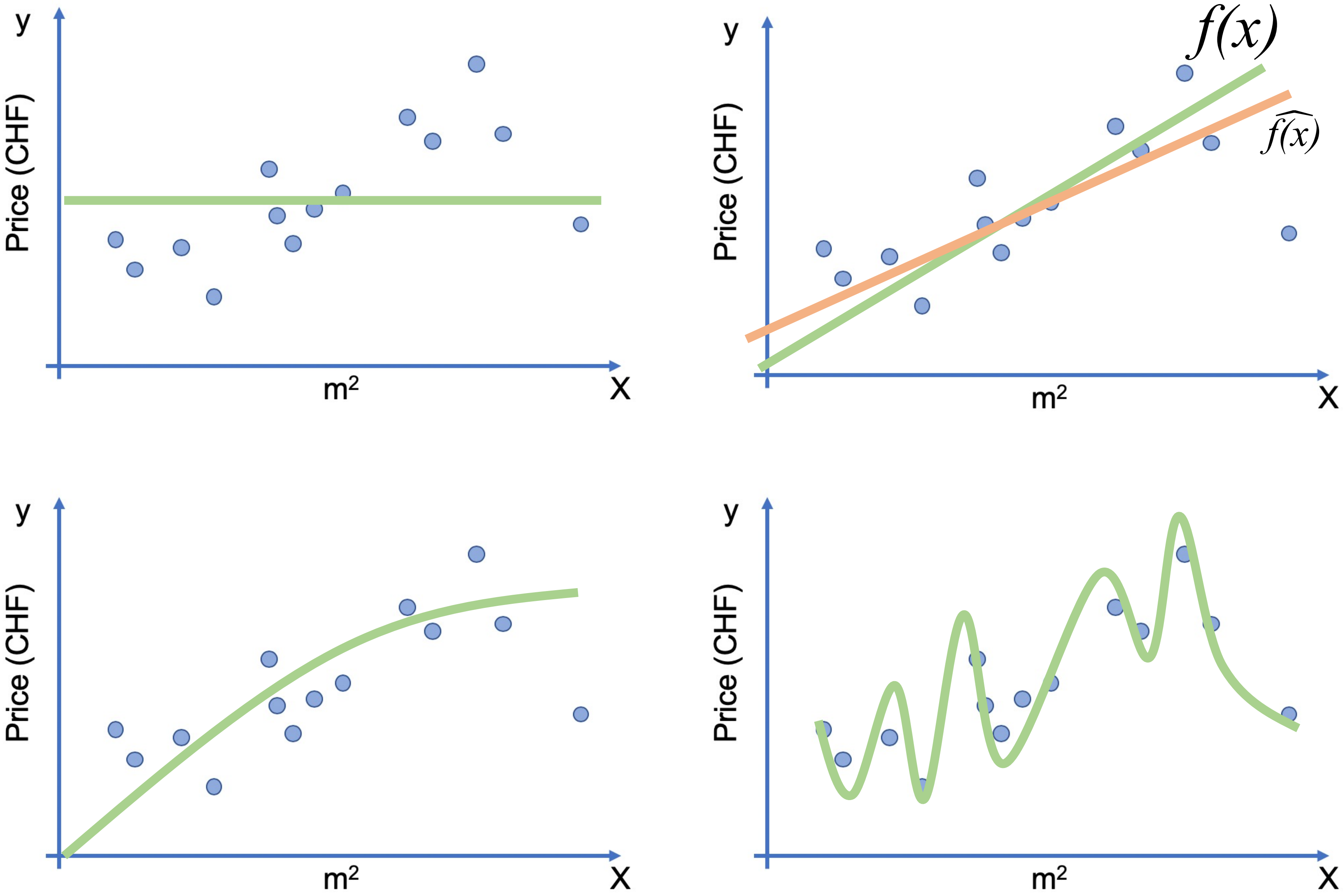
$y_i = f(x_i) + e_i$

What is a good model f(x)?

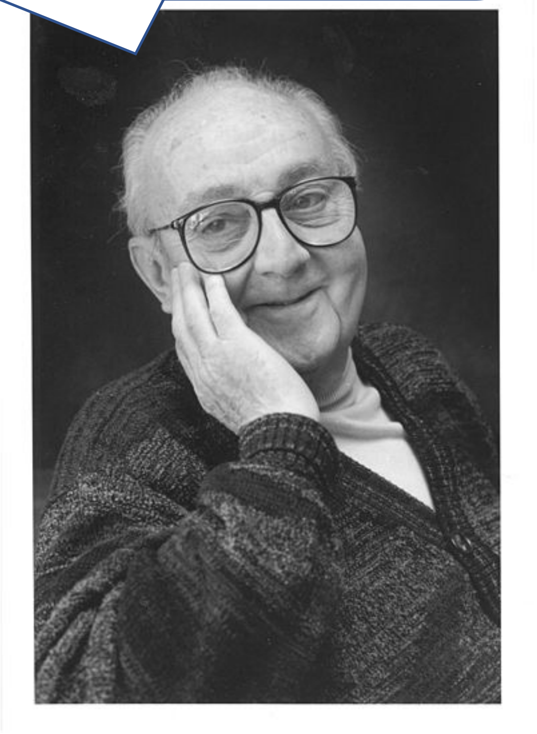


For now we will assume that there exists *linear model* $f(x)$ and we will try to approximate it with an $\widehat{f(x)}$ function which we learn from the data.

What is a good model f(x)?



All models are wrong, but some are useful!



George Box
1919 - 2013

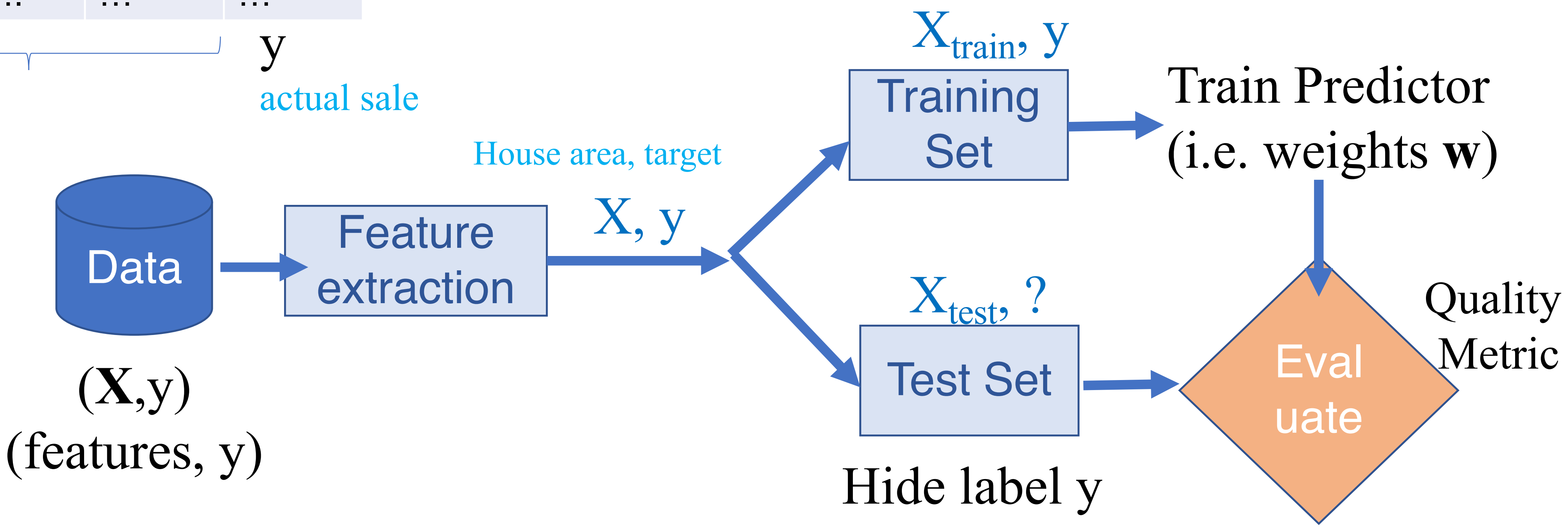
Building the predictive model

House area	Lot area	# baths	Sale Price
120	500	…	930k
65	350	…	705k
154	0	…	2010k
…	….	…	…

Data

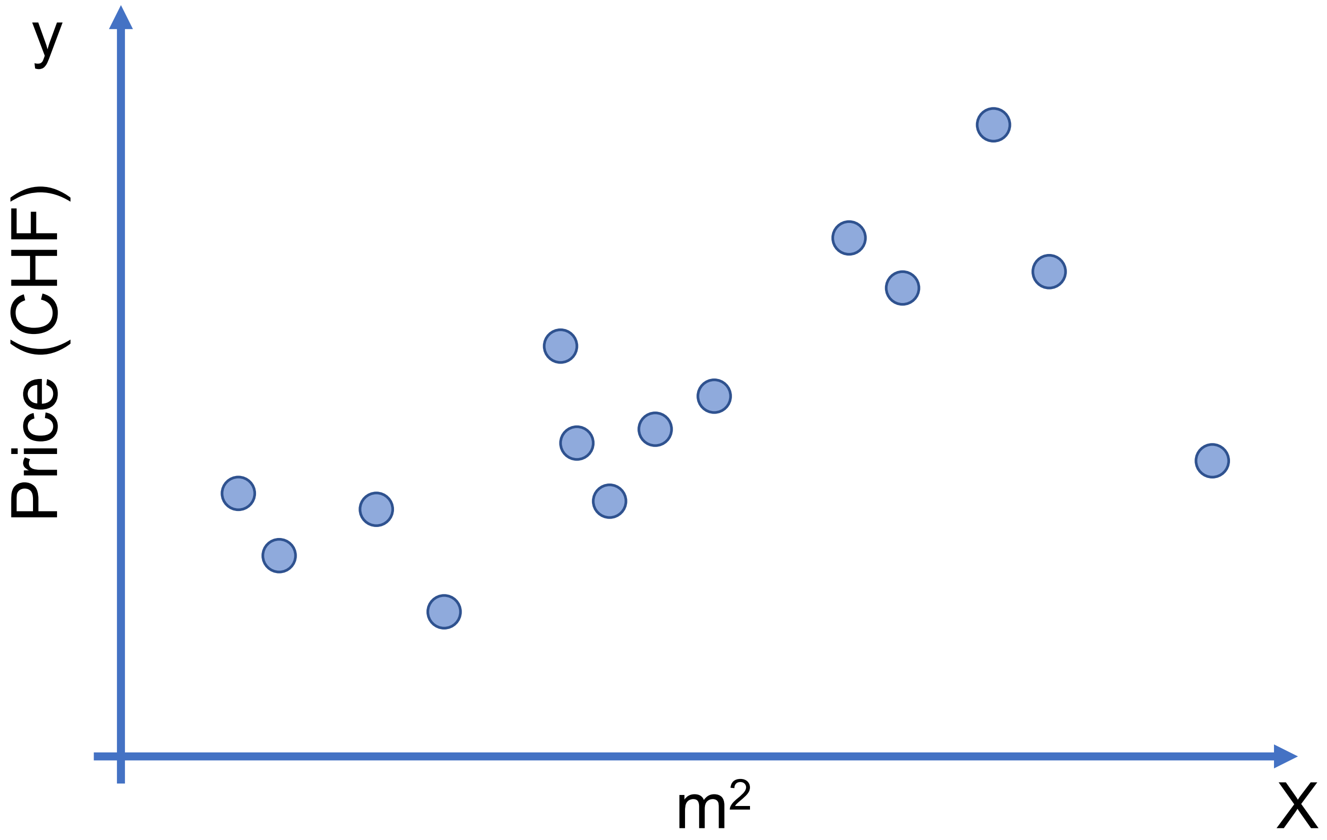
$\underbrace{\qquad\qquad\qquad}_{\mathbf{X}}$ y

actual sale



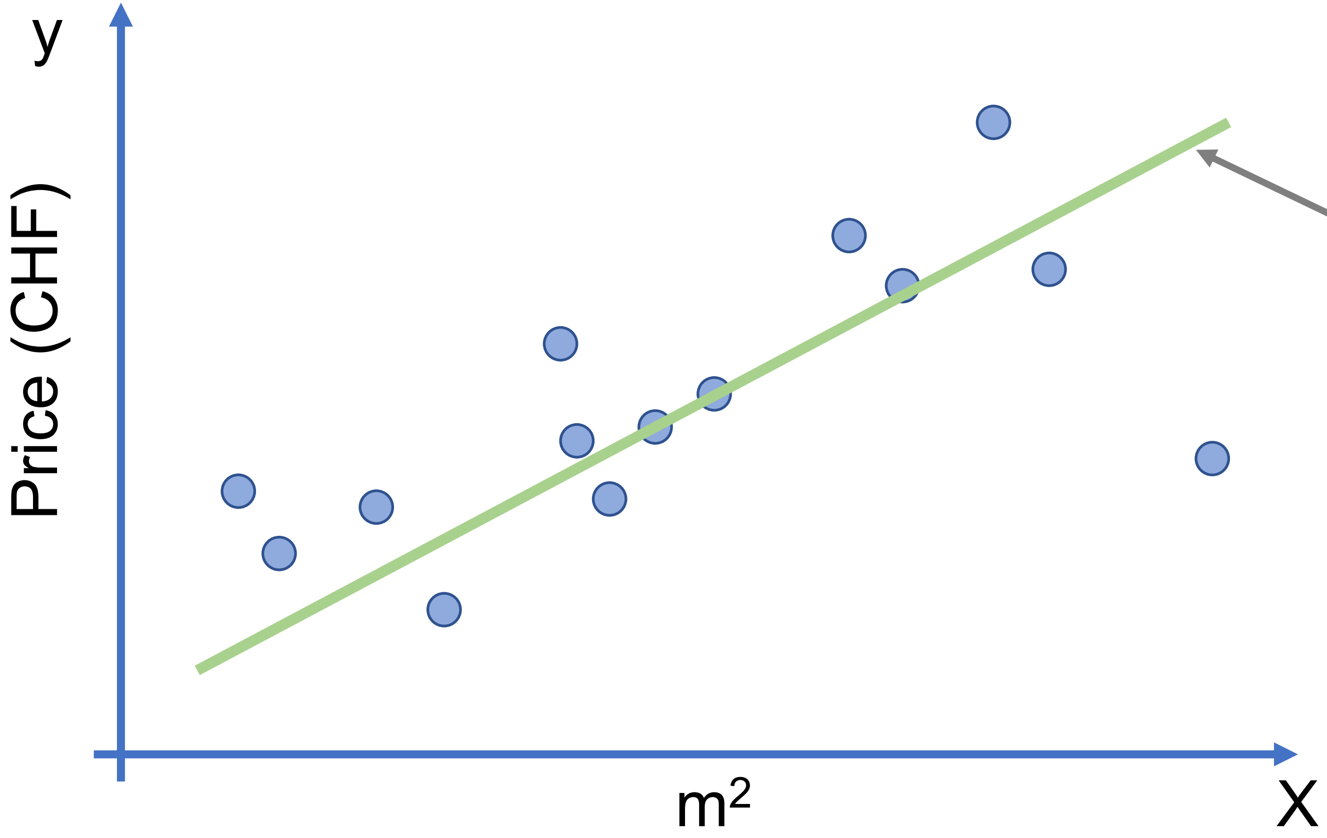
Model representation – Linear Regression

Simple linear regression model



Simple linear regression model

Fit a linear line though the data



$$\hat{y} = w_0 + w_1 x$$

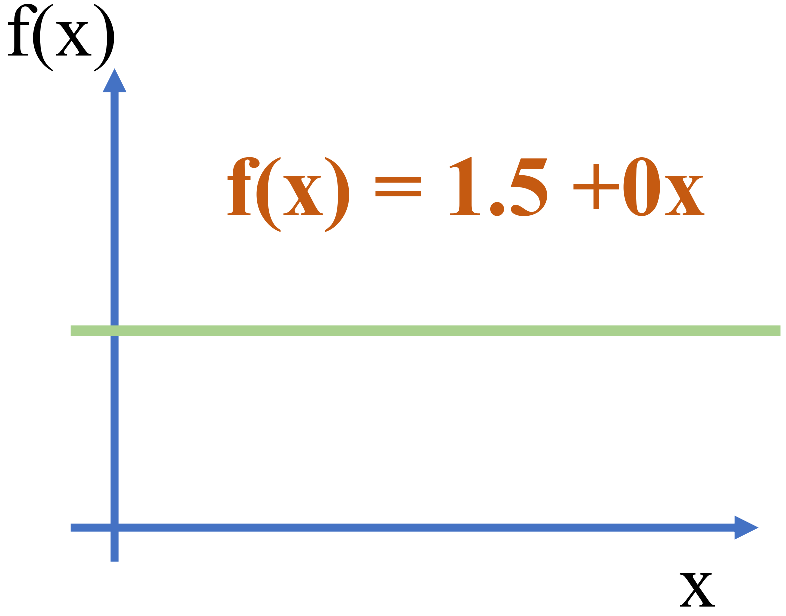
intercept

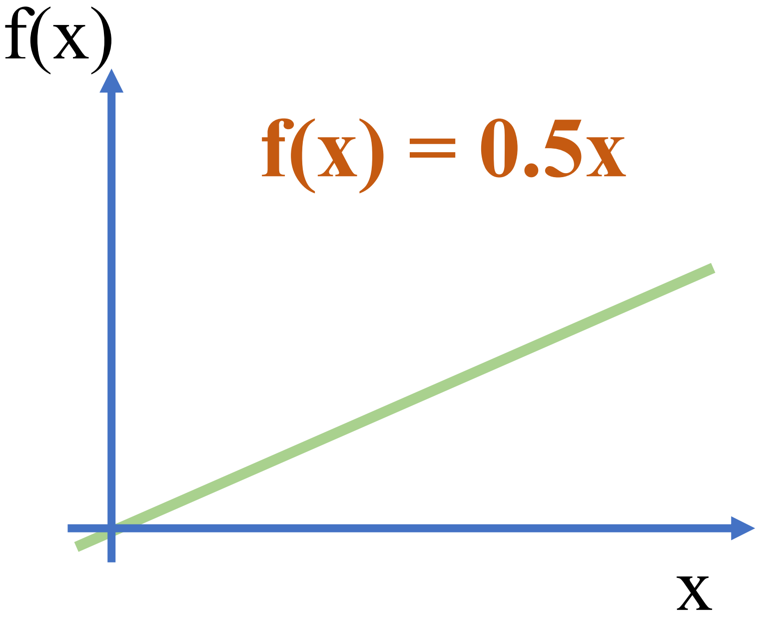
slope

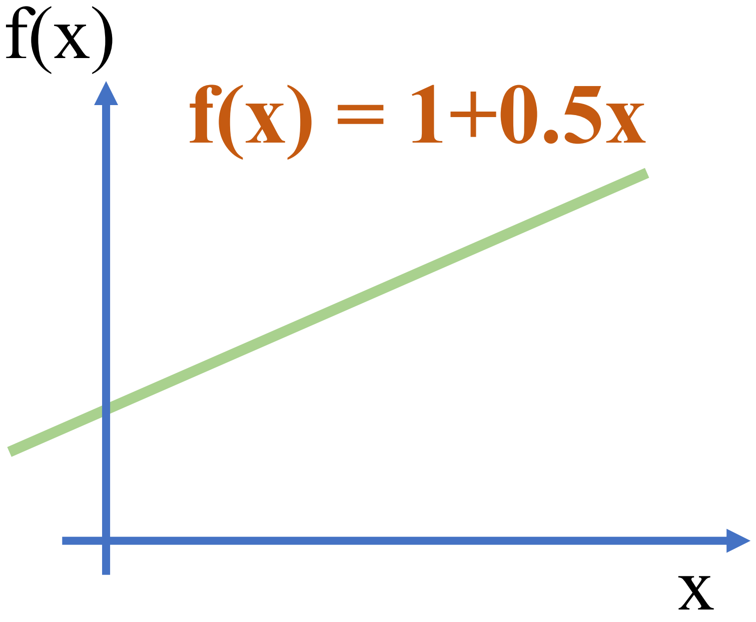
Parameters of the model
= regression coefficients

(…let's remember…)

$$f(x) = w_0 + w_1 x$$







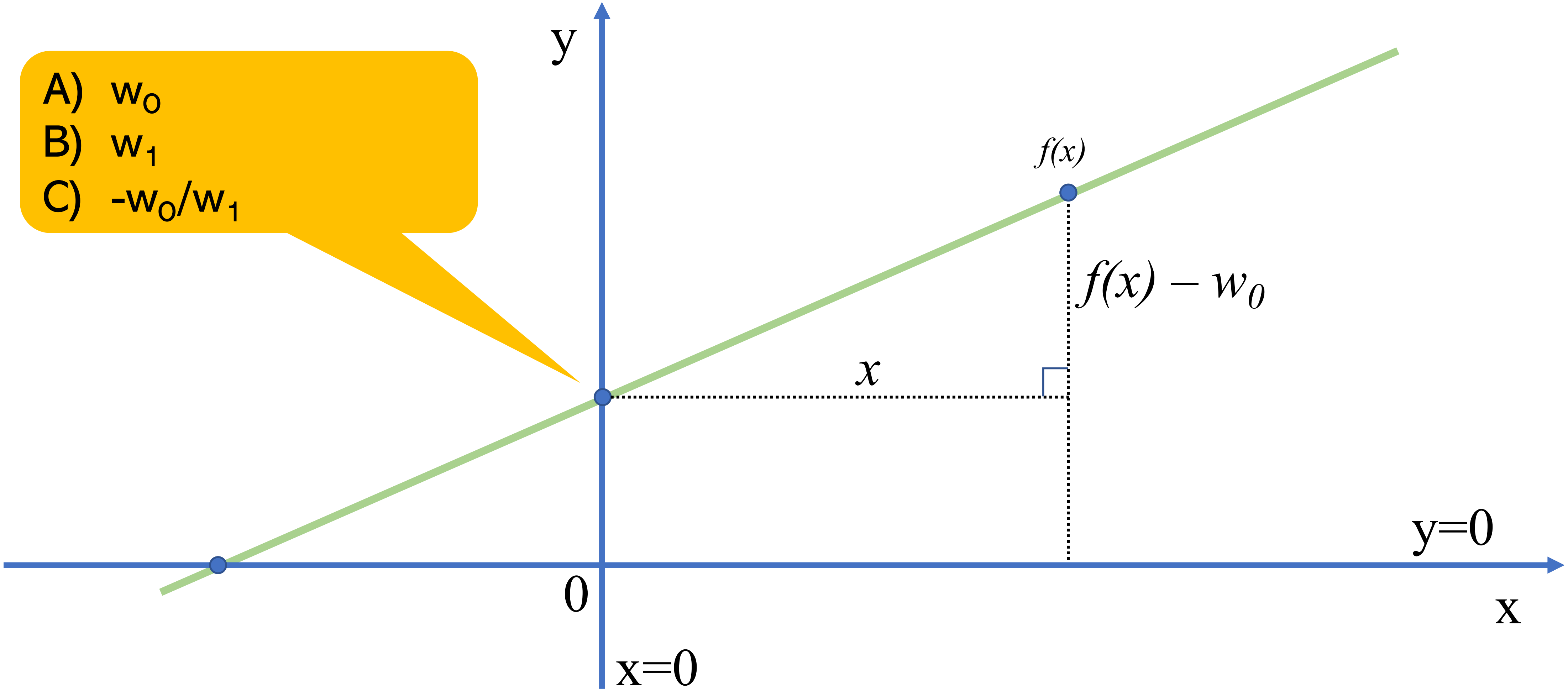
$w_0 = 1.5$
$w_1 = 0$

$w_0 = 0$
$w_1 = 0.5$

$w_0 = 1$
$w_1 = 0.5$

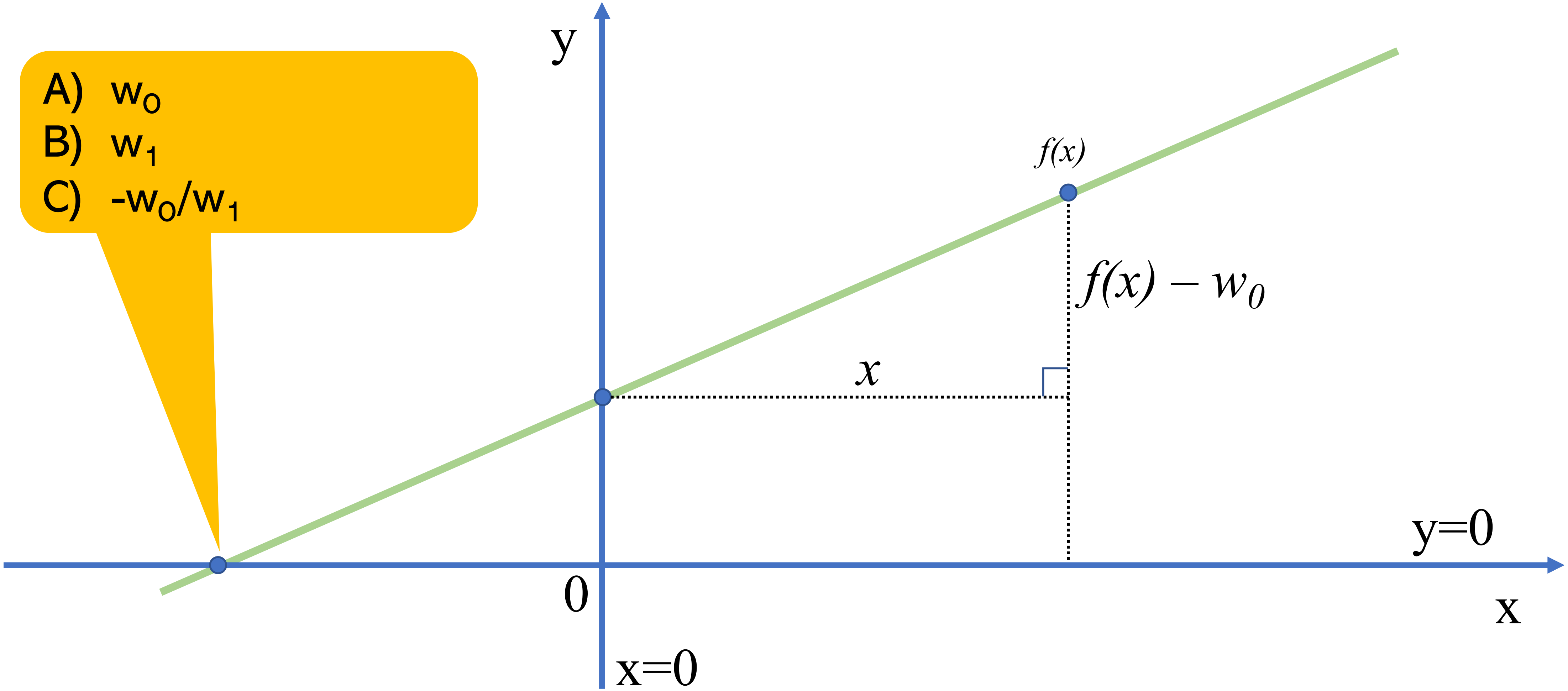
(…let's remember…)

$$f(x) = w_0 + w_1 x$$



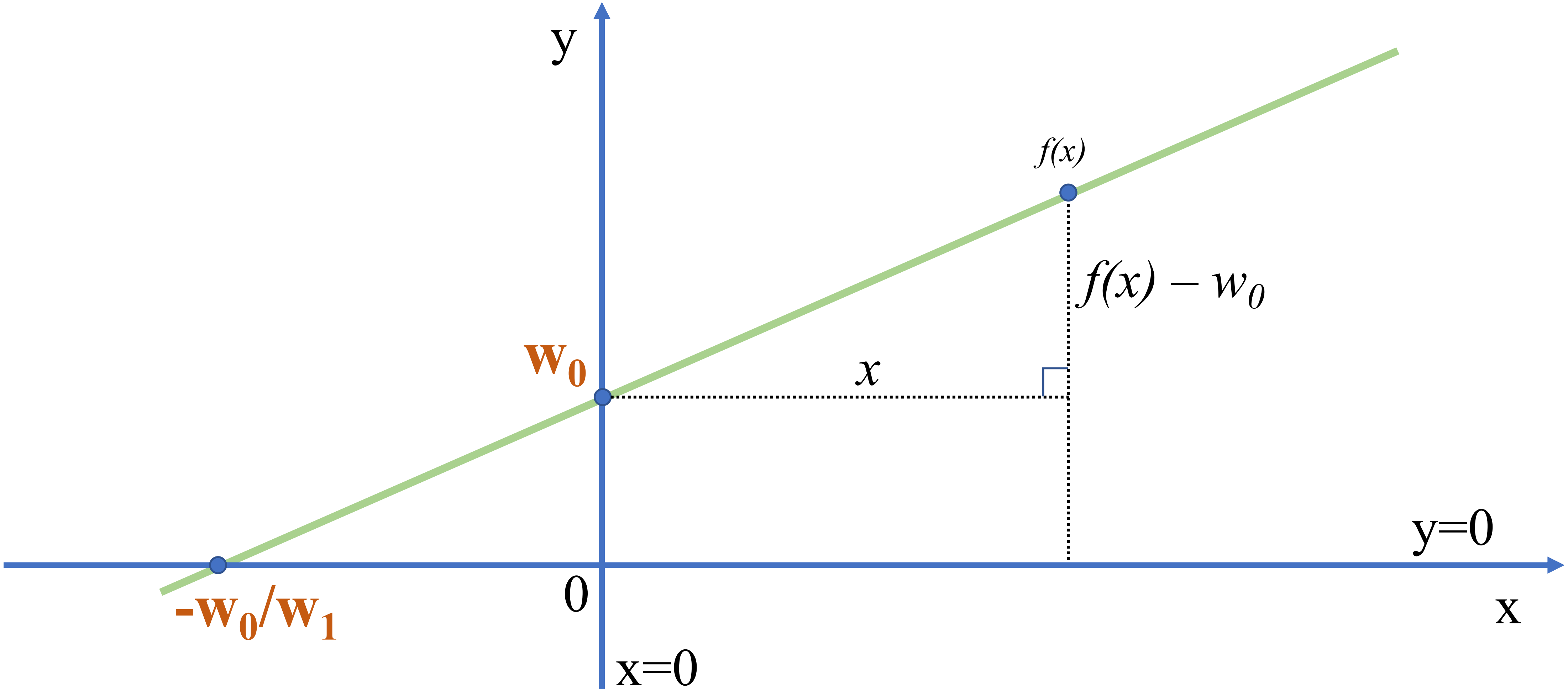
(…let's remember…)

$$\mathbf{f(x) = w_0 + w_1 x}$$



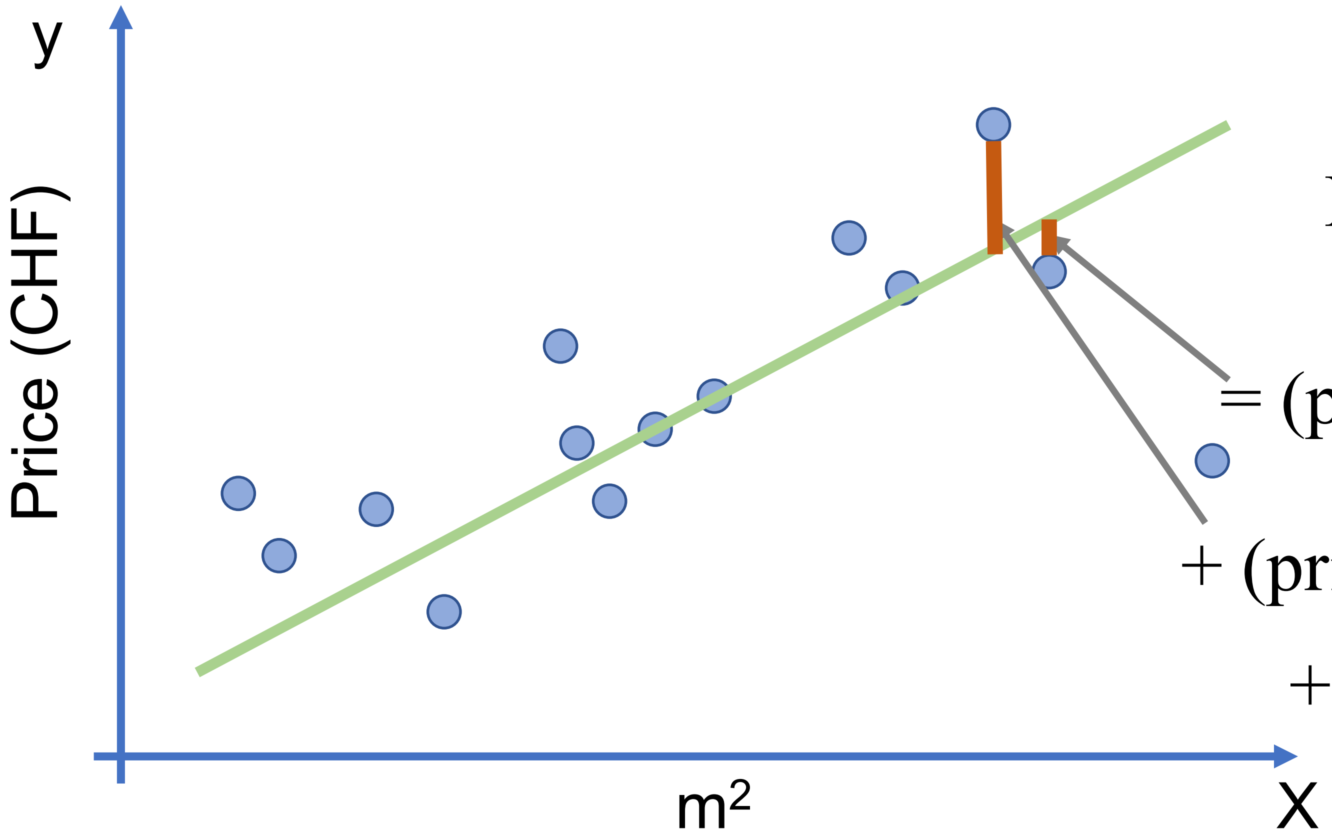
(…let's remember…)

$$f(x) = w_0 + w_1 x$$



Error of using some line model

Residual sum of squares (RSS)



actual value

predicted value

$$\mathrm{RSS}(w_0, w_1) = \sum_{i=1}^{N} (y_i - (w_0 + w_1 x_i))^2$$

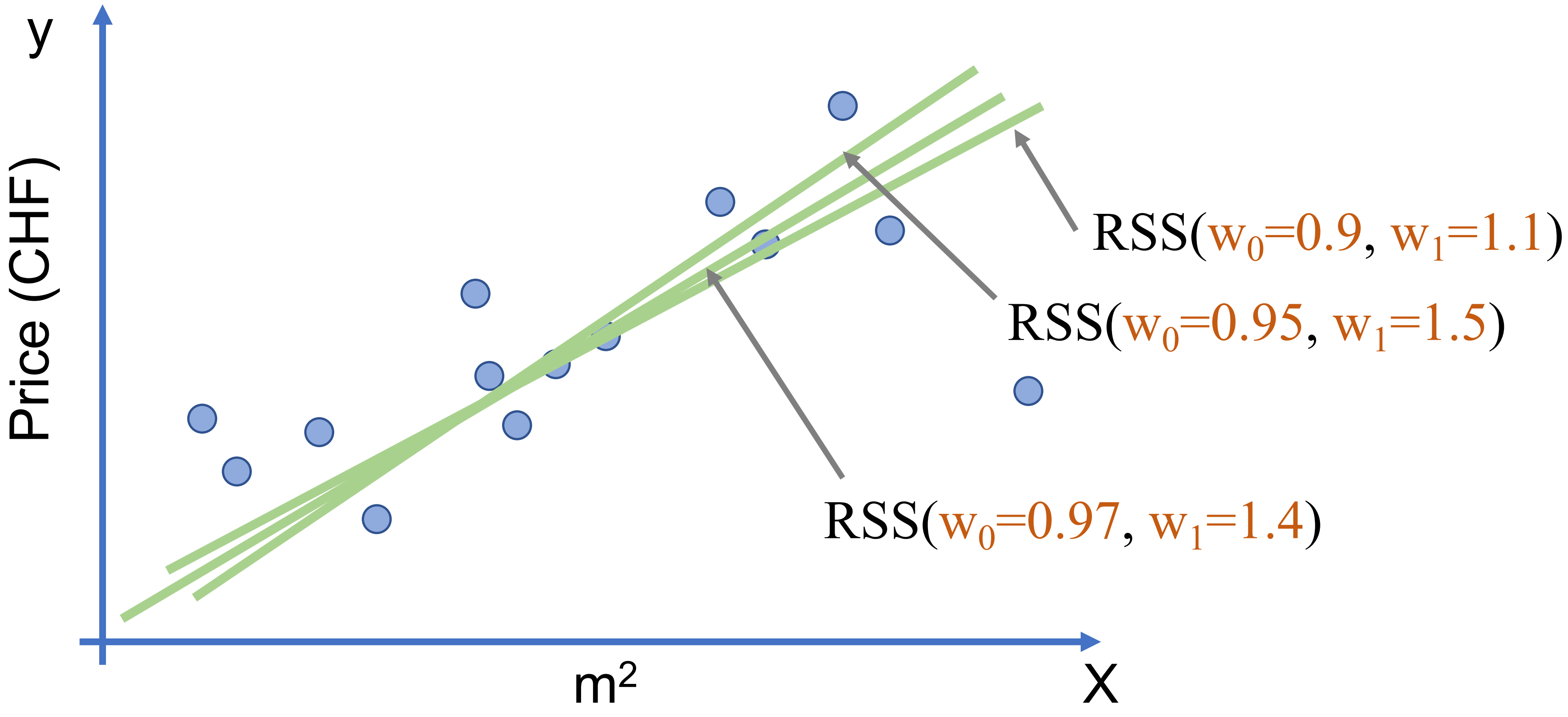
$$= (\text{price\_house1} - (w_0 + w_1 \text{house\_area1})^2$$

$$+ (\text{price\_house\_2} - (w_0 + w_1 \text{house\_area2})^2$$

$$+ \ldots$$

Find the best line

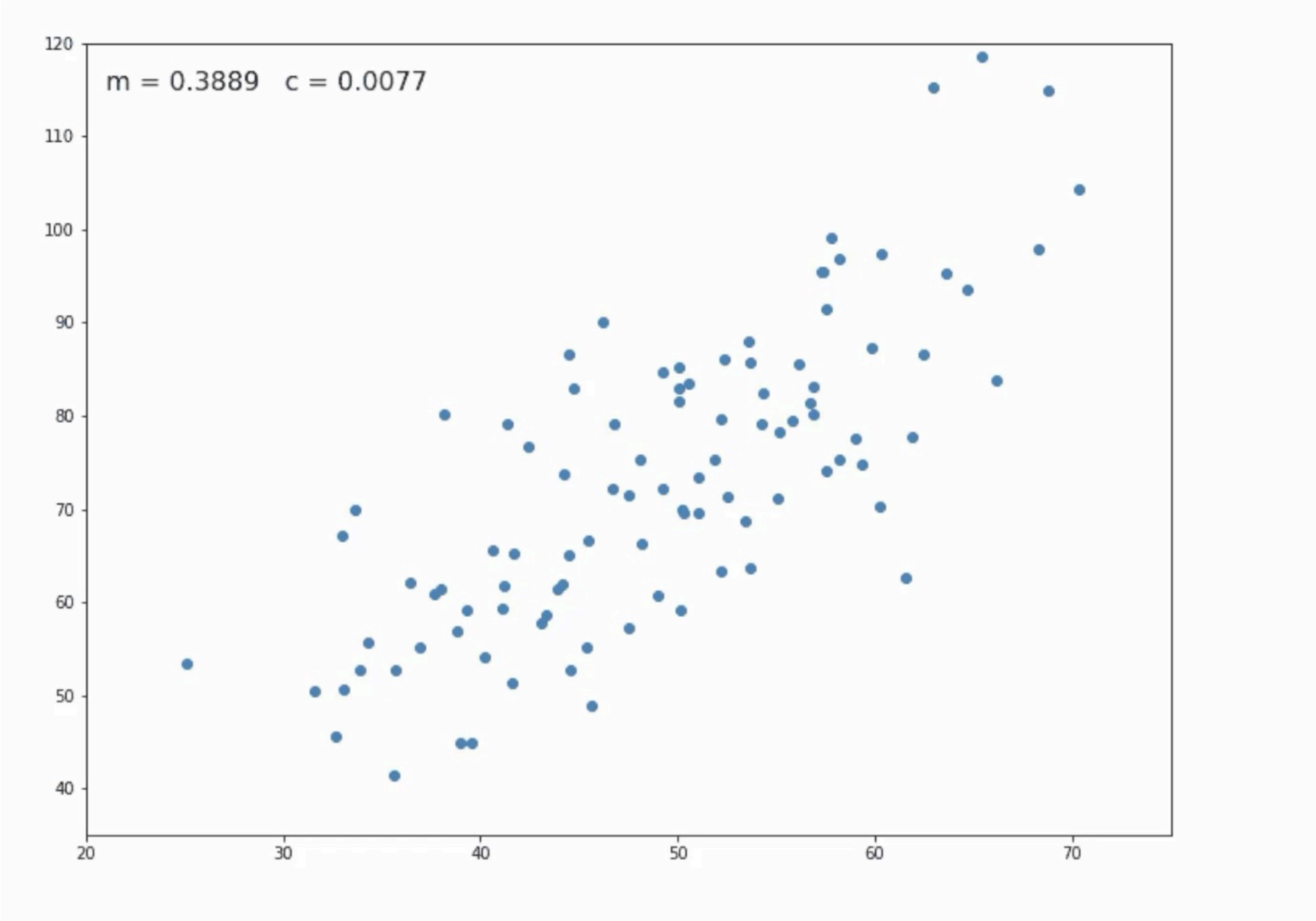
Minimize the cost across all possible w_0, w_1



How to find the best line – Gradient Descent

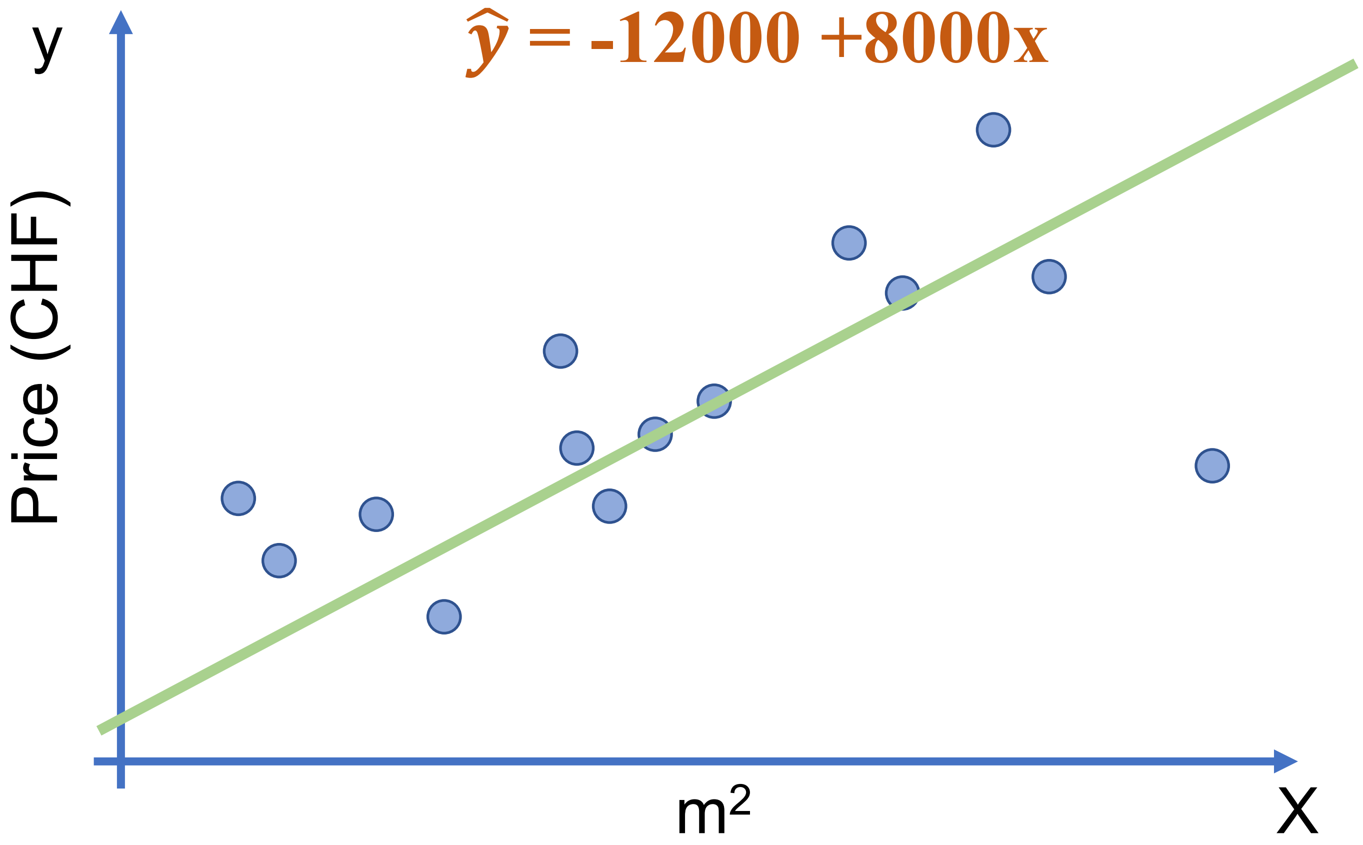
- Typically, we find the best line using a method called **"gradient descent"**. This is an iterative method that tries different solutions and guides the tries based on the error (RSS in this case).
- Gradient descent does not try *randomly* for different solutions, but tries values that look most promising and tries in every step to improve the previous estimate.

Demo of gradient descent



Doing predictions

Assume this is the model with the lowest RSS cost

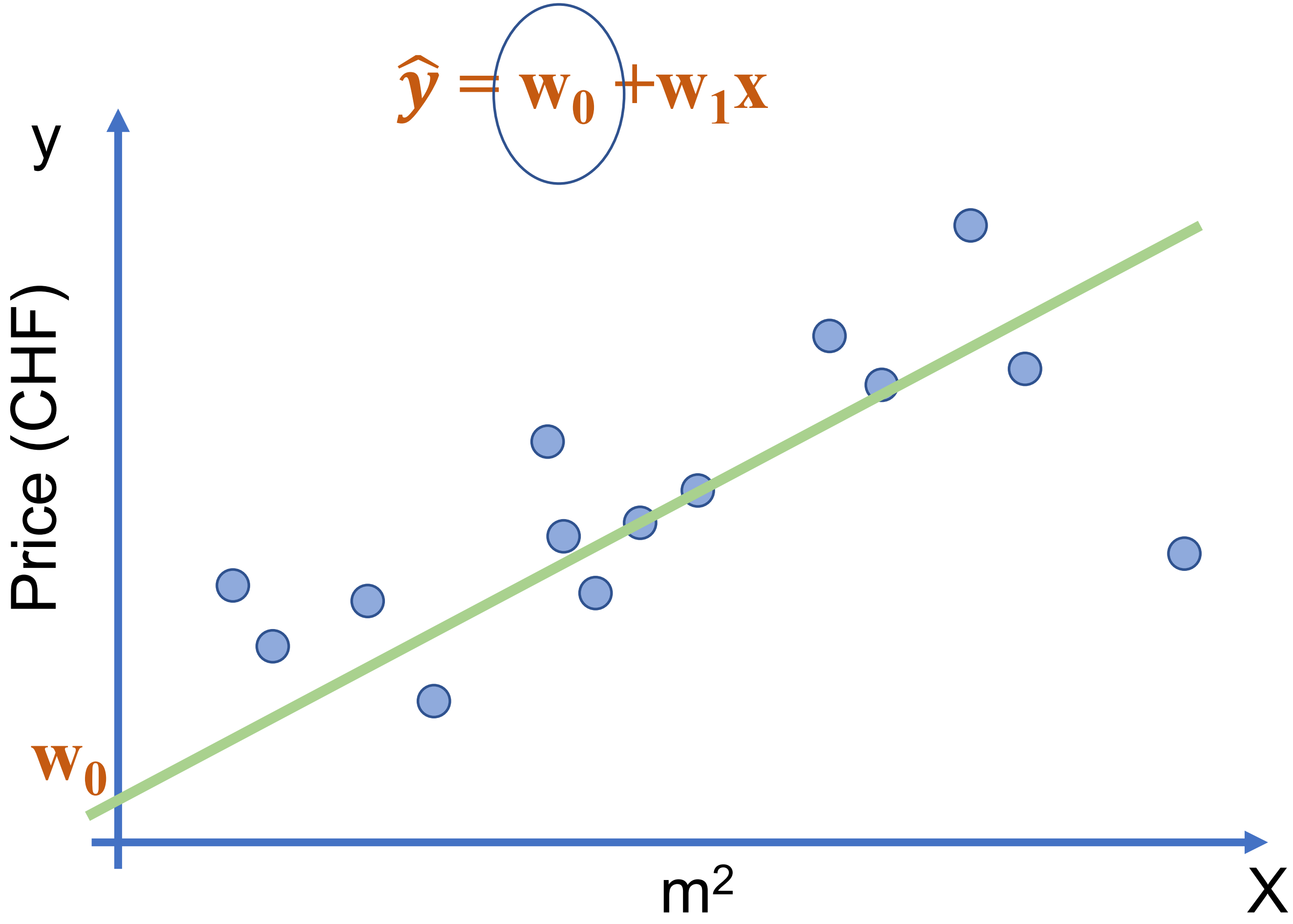


What is the price for a house with area of 100 sq m^2 ?

Price = -12000 +8000*100
= 788'000 CHF

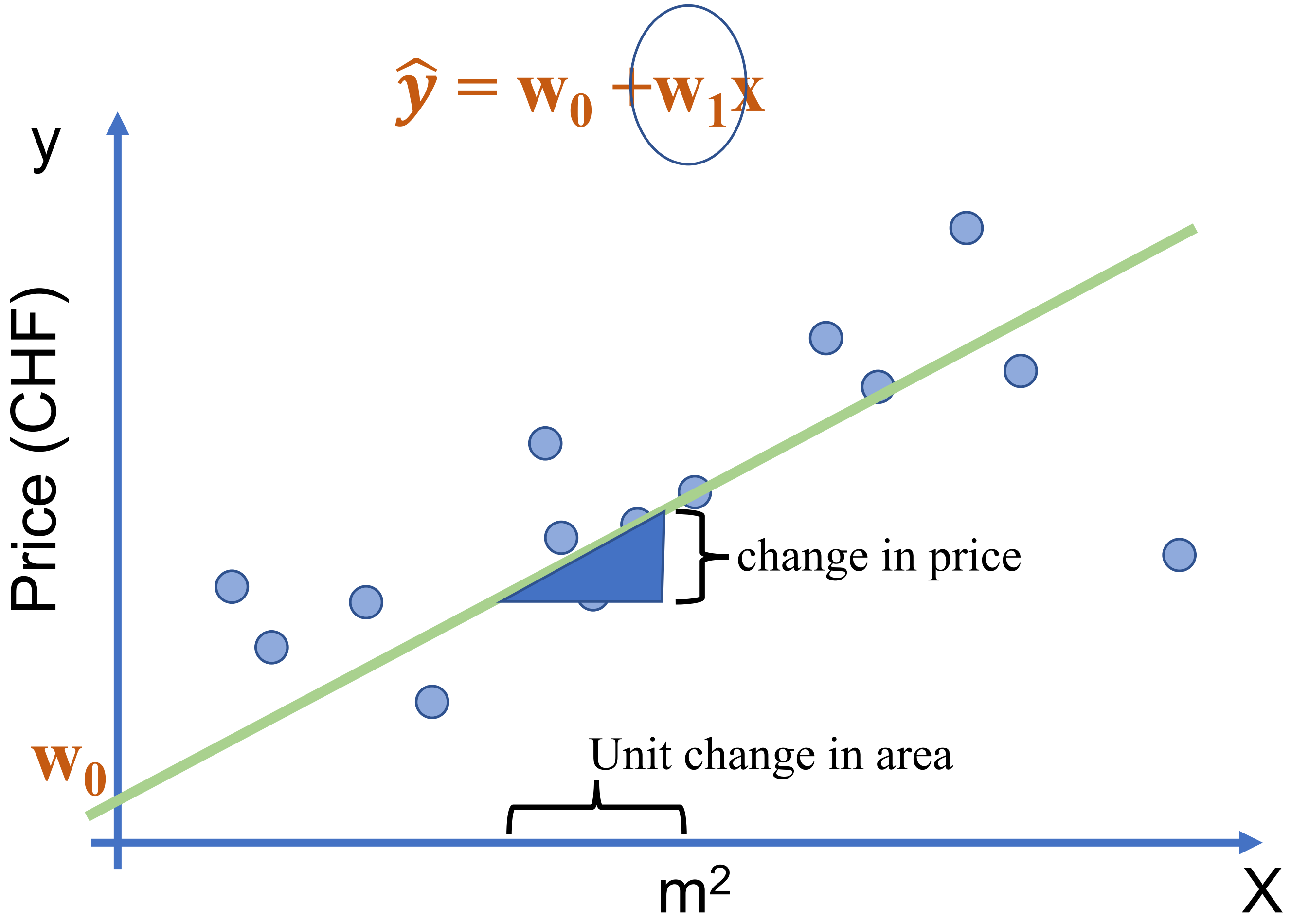
What is the area for a house sold for 1'200'000?

Interpreting the coefficients



Predicted price for house with zero area (not very meaningful in this case)

Interpreting the coefficients

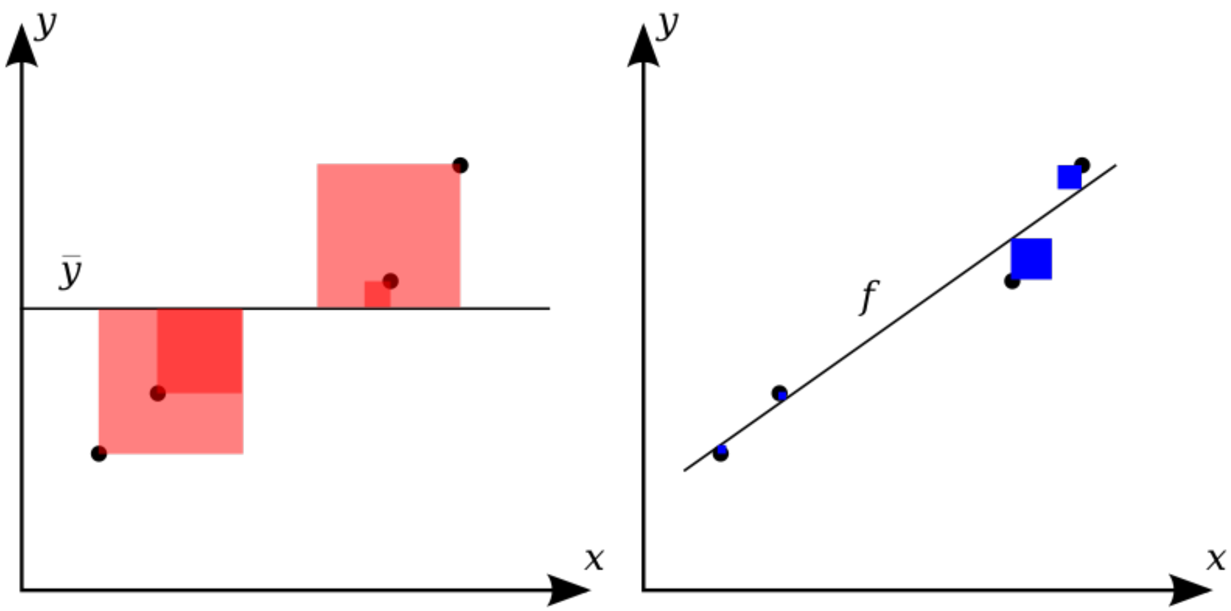


Predicted change in house price for 1 unit change in house area.

Exercise:
Compute
$\hat{y}(101) - \hat{y}(100)$

Goodness of fit (R^2)

- Typically in statistical analysis we use a number such as R^2 which encodes how much of the data variance is explained by the model.
- $R^2=1$ (perfect model)
- $R^2=0$ (same as baseline, ie avg)



$$R^2 = 1 - \frac{RSS}{TSS}$$

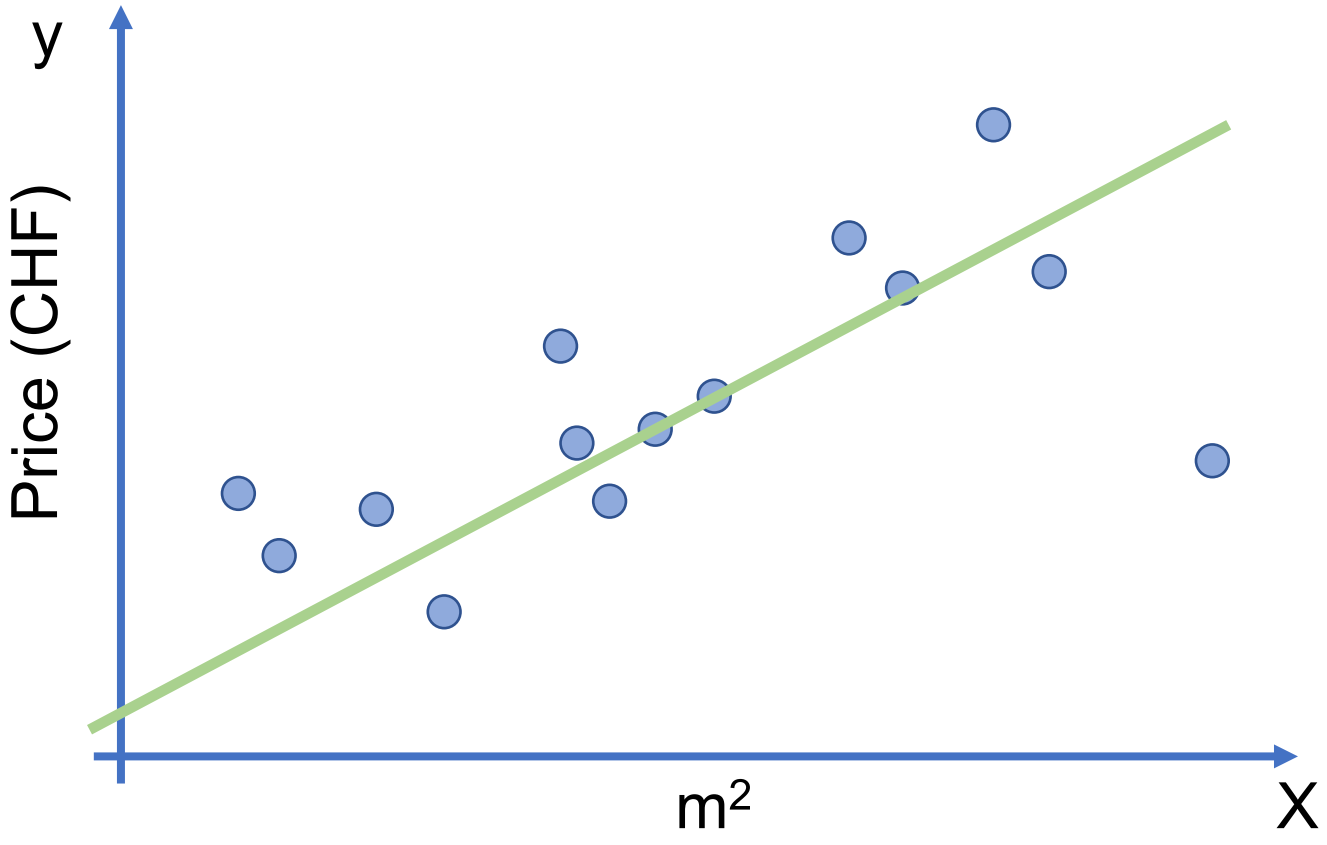
Explained variance

Total variance

Residual Sums of Squares: $\mathrm{RSS} = \sum(y_i - \hat{y}_i)^2$

Total Sums of Squares: $\mathrm{TSS} = \sum(y_i - \bar{y})^2$

Goodness of fit (MAE)



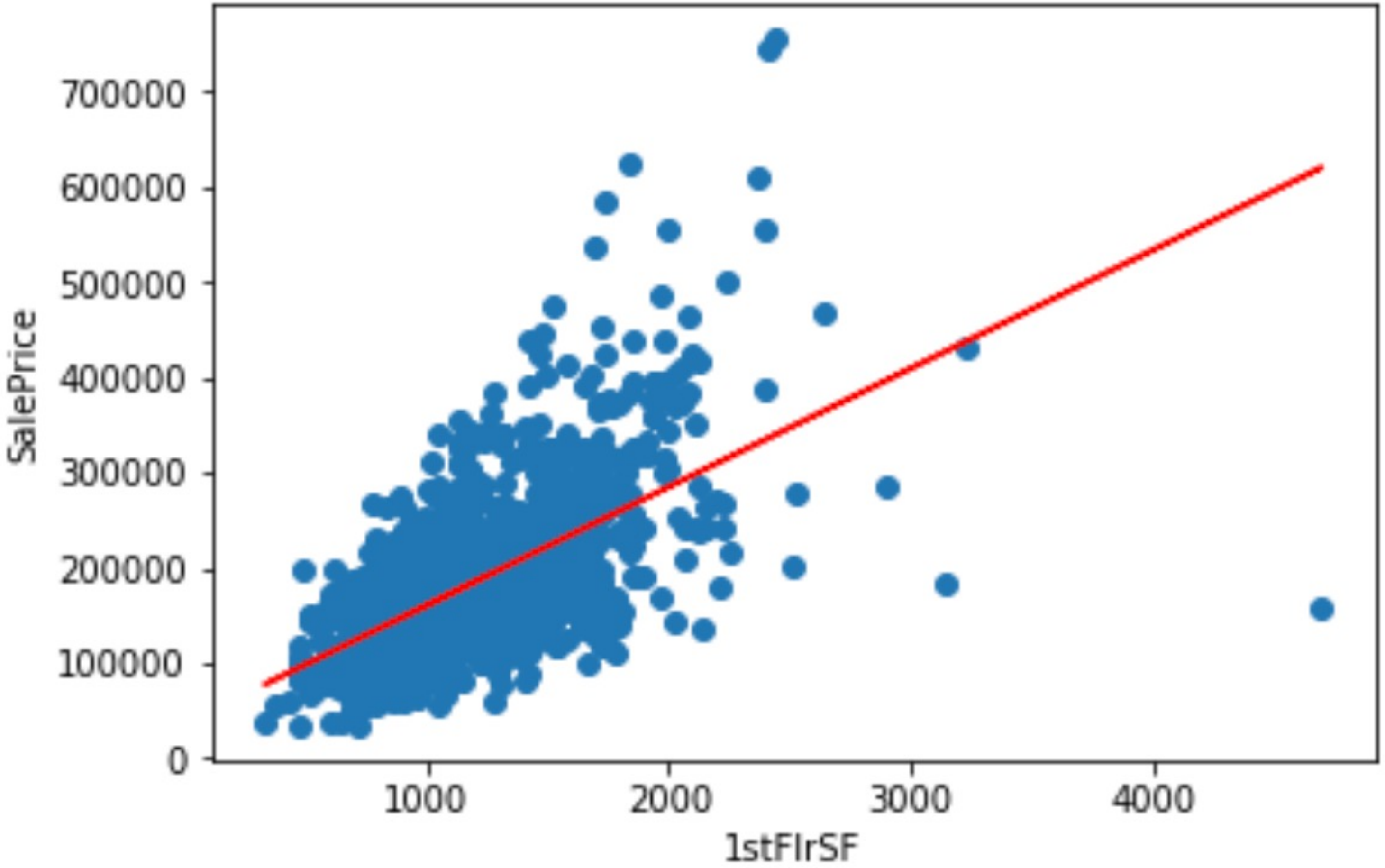
MAE(Mean Average Error):

$$1/N \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Can provide some more meaningful number (compared to RSS)
Eg if MAE = 25’000 it means that *on average* we are off by 25k CHF in our predictions.

In-class demo – Predicting house prices

https://tinyurl.com/DMML-regression



In-class Exercise (5 mins)

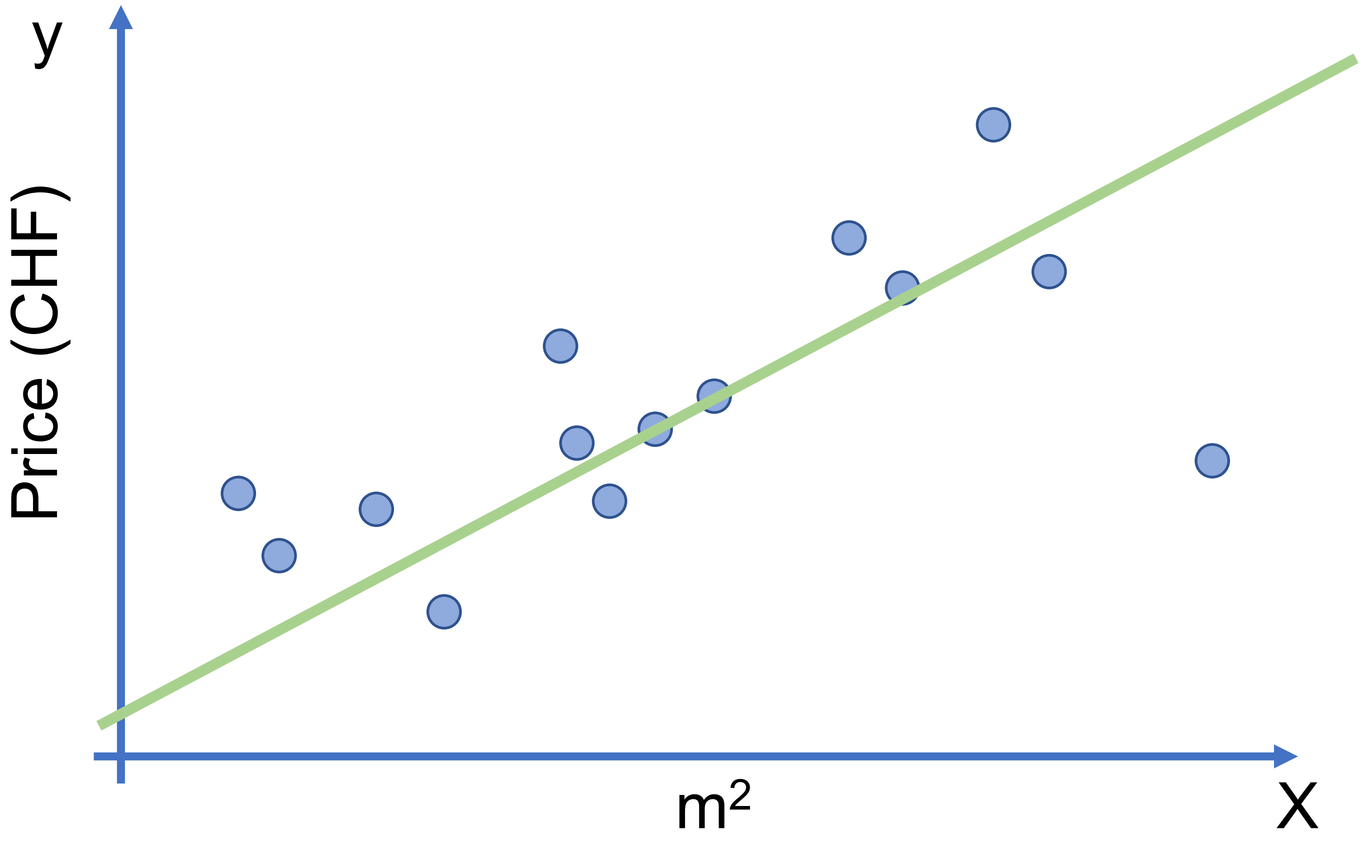
- Do a regression of the features
  - `FullBath` vs `SalePrice`
  - `OverallQual` vs `SalePrice`

Which has the lowest Mean Average Error (MAE) and R^2?

Multiple Regression

Adding more features

So far we made predictions using only the house size

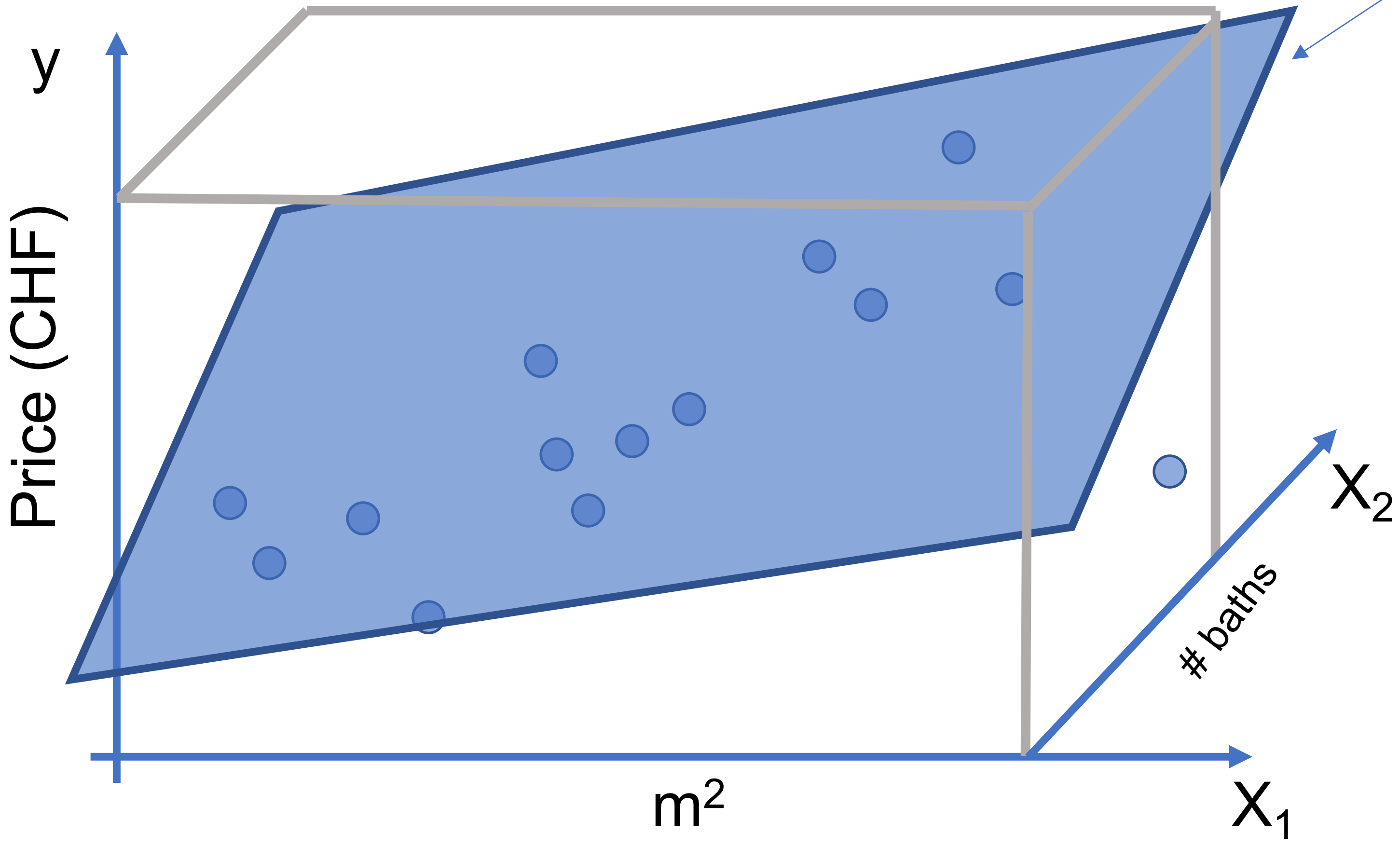


Adding more inputs

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2$$

m²

#baths



Many possible inputs

- House area
- Lot area
- # bathroom
- # bedrooms
- Year built
- …

Examples: n = 4; dimensions: d = 4

x_0	House Size x_1	# Rooms x_2	# Bathrooms x_3	Landsize x_4	Price y
1	120	4	2	200	900’000
1	132	5	1	15	1’200’000
1	98	2	1	0	850’000
1	85	2	1	49	970’000

$$X = \begin{bmatrix} 1 & 120 & 4 & 2 & 200 \\ 1 & 132 & 5 & 1 & 15 \\ 1 & 98 & 2 & 1 & 0 \\ 1 & 85 & 2 & 1 & 49 \end{bmatrix}$$

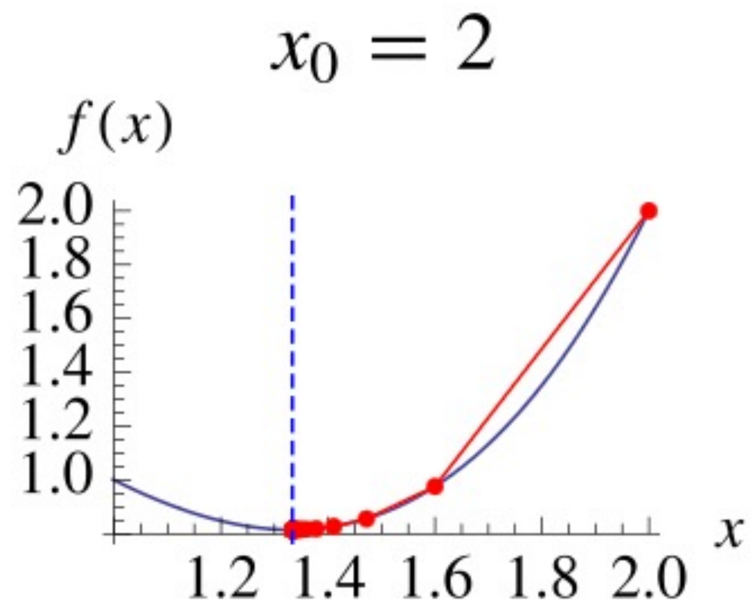
n x (d+1) array

$$y = \begin{bmatrix} 900000 \\ 1200000 \\ 850000 \\ 970000 \end{bmatrix}$$

n-dimensional vector

Steepest (gradient) descent

- To find the weights of the linear regression (or of most ML problems) we usually use an approach called gradient descent.



See also here from a demo:
https://demonstrations.wolfram.com/CurvesOfSteepestDescentFor3DFunctions/

For those interested in more details, have a look at the advanced (optional) slides.

More regression examples

Can you come up with some more??

Work in Two:

- [3mins] Work with your neighbor.
- Find some more interesting regression problems.
  - How would you model it as a regression?
  - What are the features that influence the target variable?

- [3mins] We discuss.

Determine salary

- How much salary y will you get after you graduate?
- Depends on X=
  - total grade,
  - #internships,
  - years of past experience,
  - #friends on Facebook,
  - #people you connect with in LinkedIn,
  - …

$$\hat{y} = w_0 + w_1 grade + \mathrm{w}_2 experience + \mathrm{w}_3 Facebook\_friends + \ldots$$

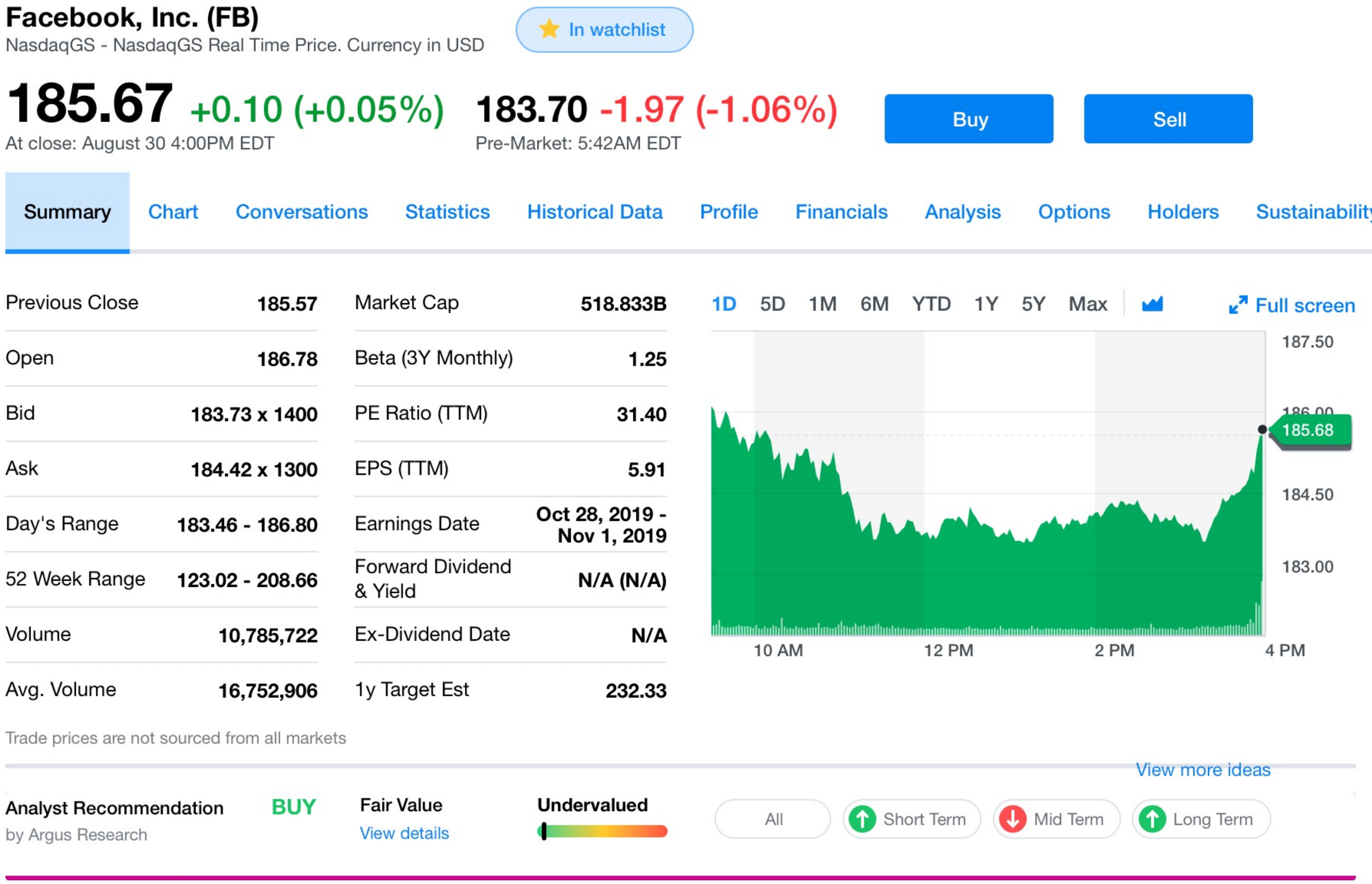
weight

but cannot compare the weight unless i normalize the data.

Stock price prediction

Depends on:

- Last day's *value*
- Last week's price *slope*
- How many events in the news about the company
- # times Donald Trump mentioned the company
- …

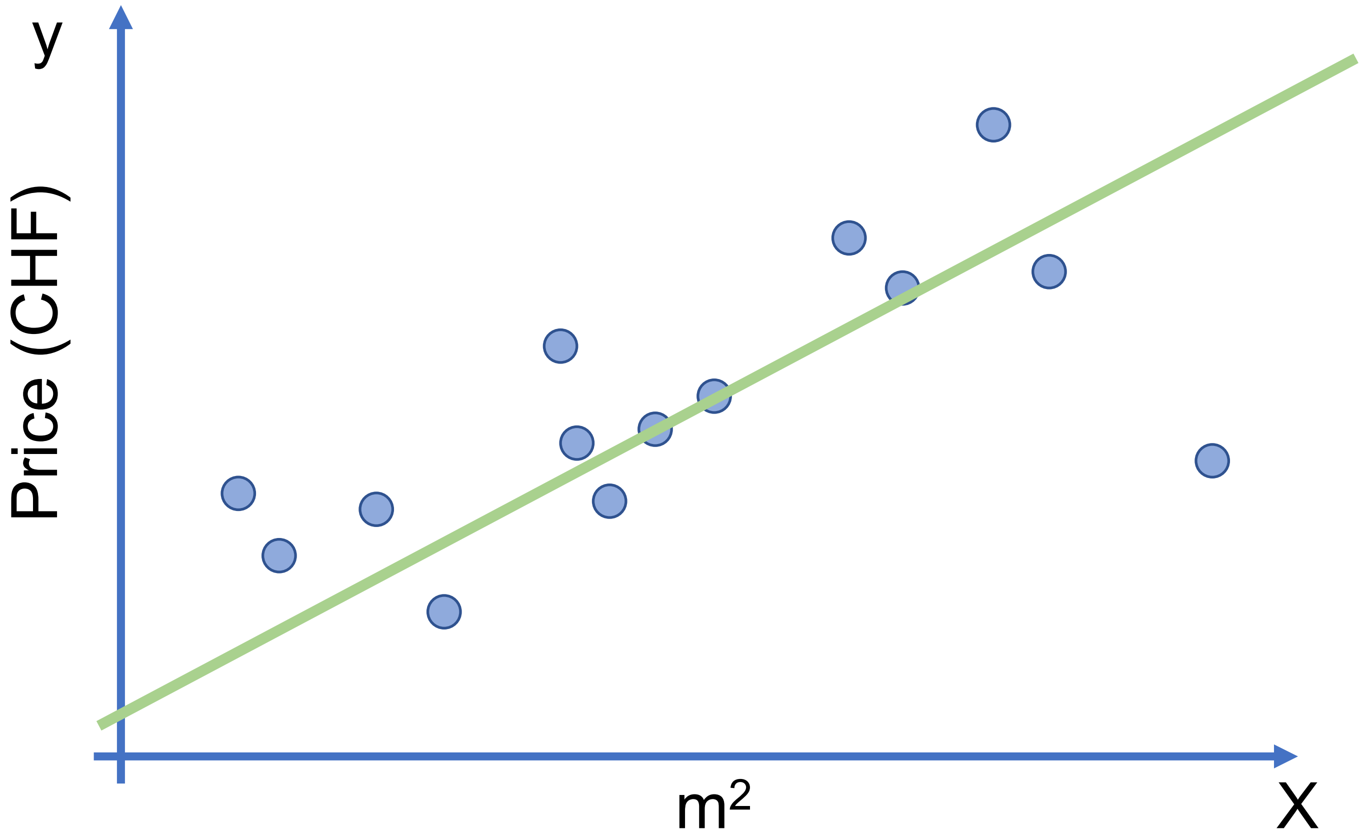


Predict number of retweets

- How many people will retweet your post?
- Depends on:
  - # followers
  - # hashtags in post
  - Popularity of hashtags
  - Text in post
  - # images in post
  - …

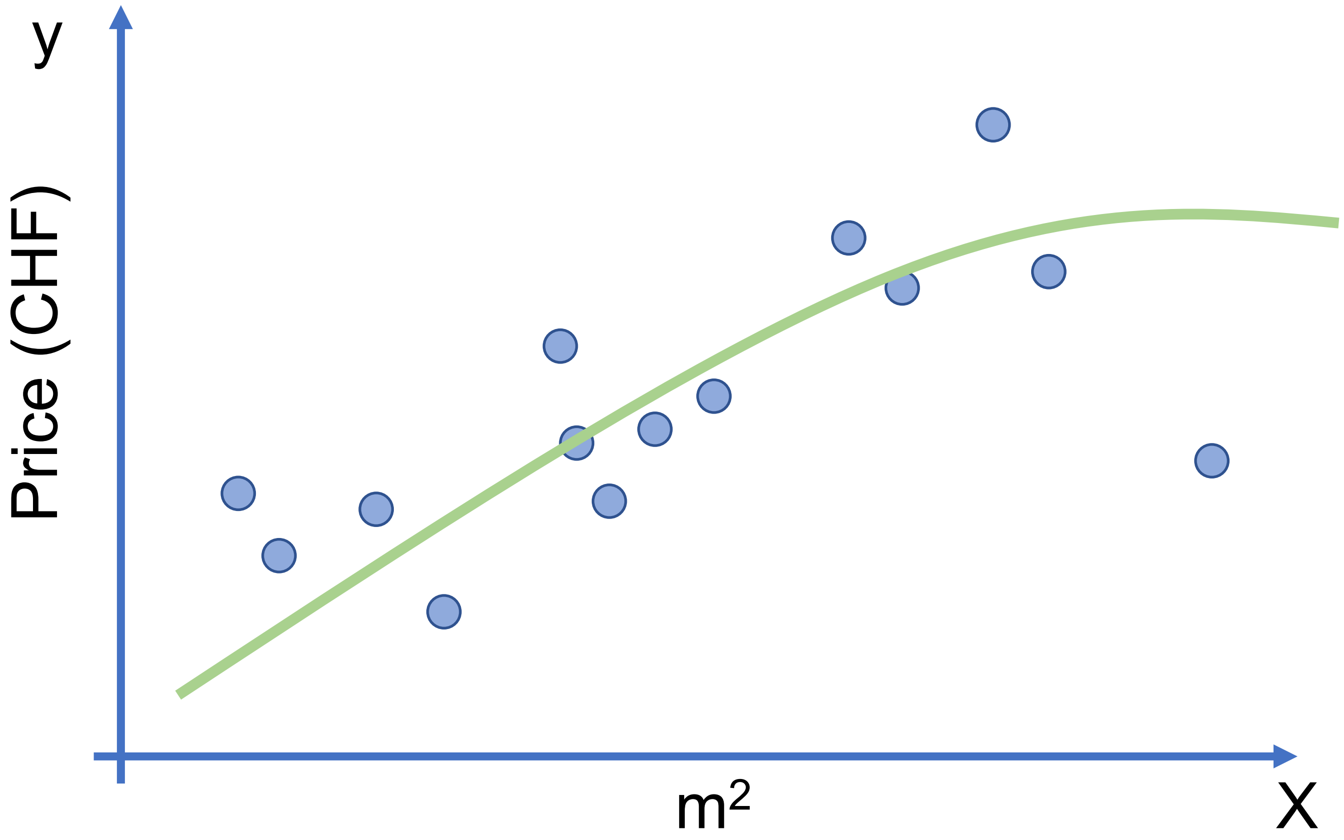
Building more complex models

Our linear model



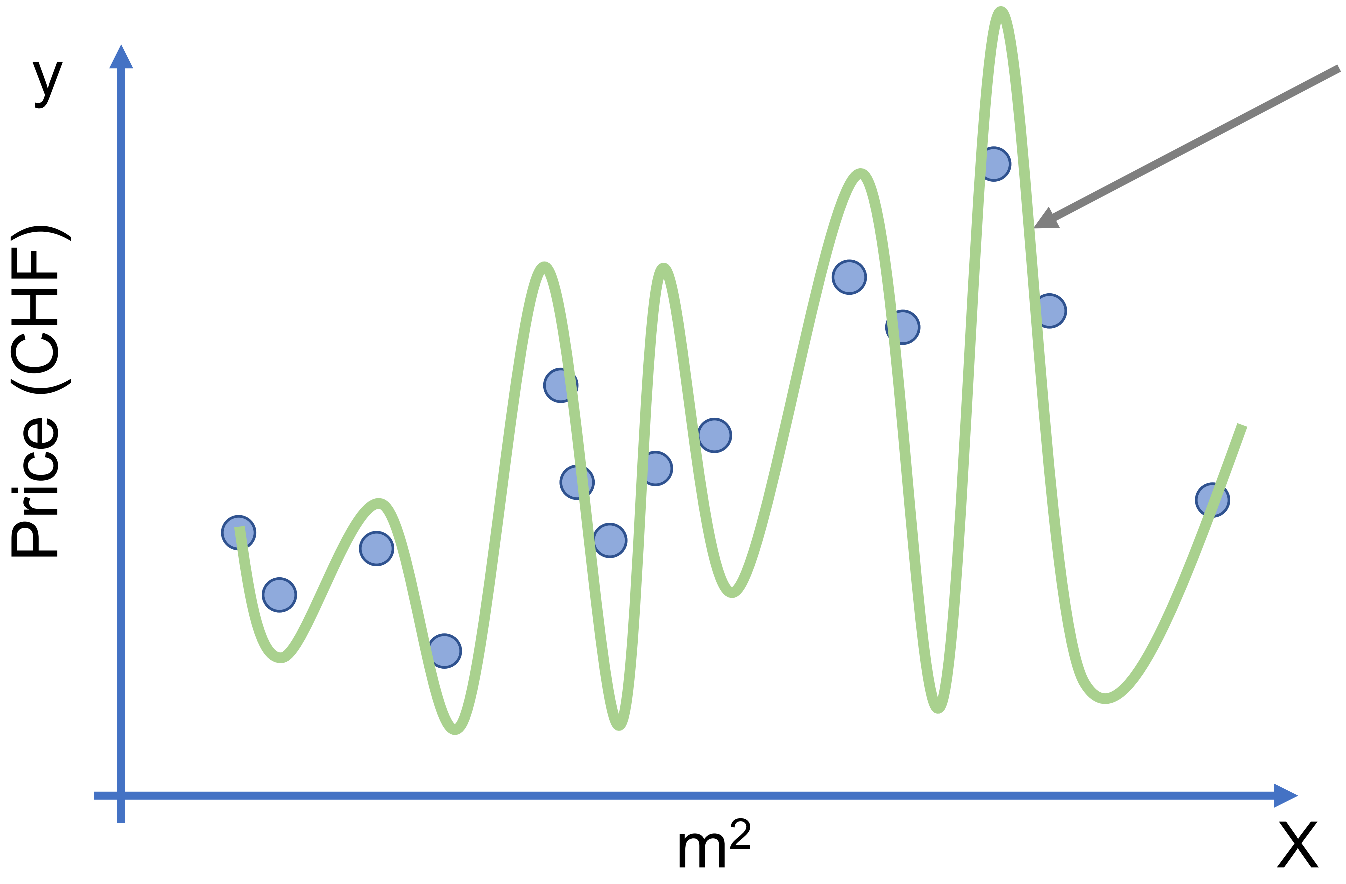
RSS = 130

A quadratic model



RSS = 110

A higher-order polynomial



$$f(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_p x^p$$

$\boxed{RSS = 15}$

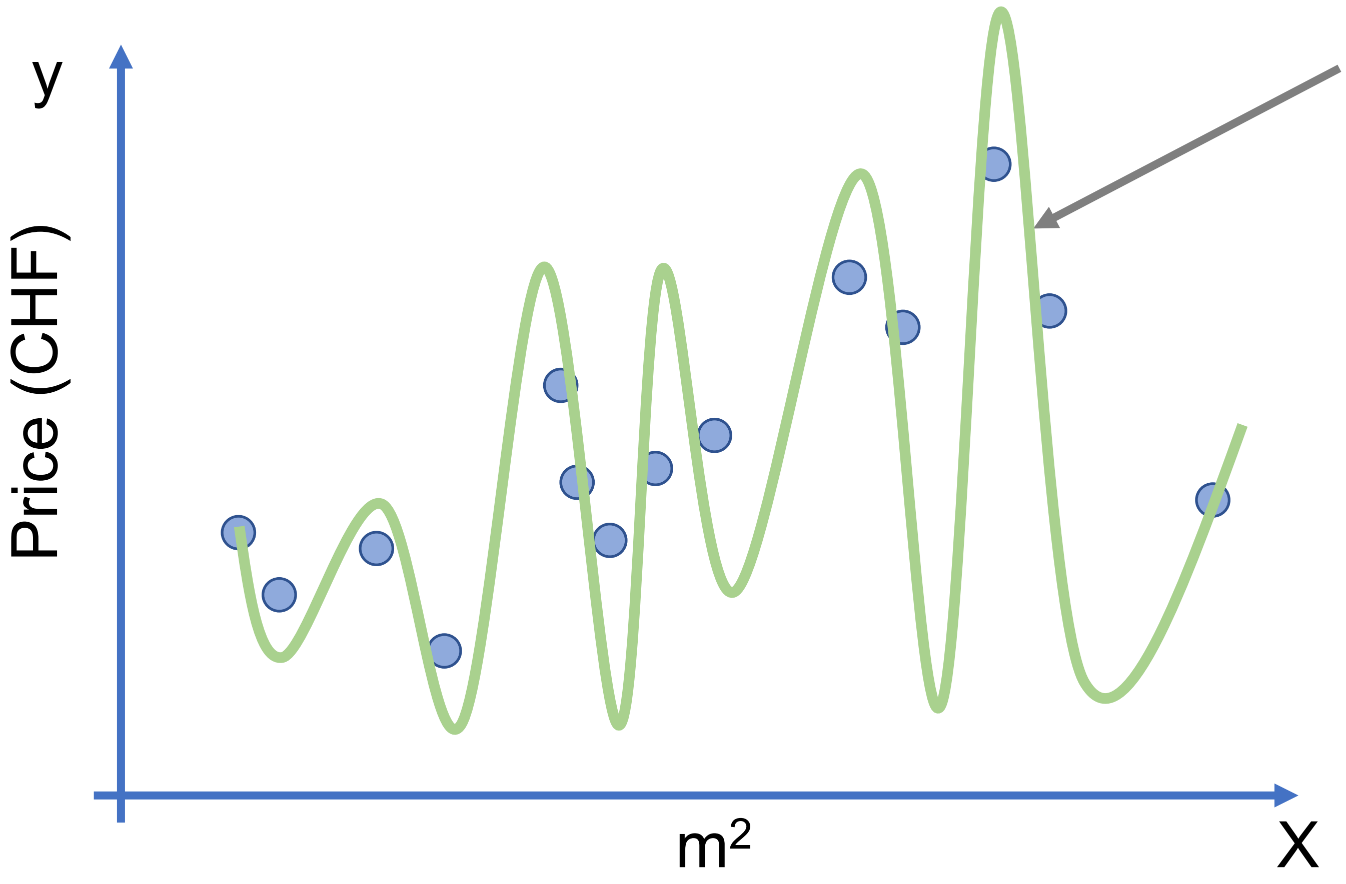
Example: Polynomial regression

x_0	House Size x	House Size² x^2	Price y
1	120	120^2	900'000
1	132	132^2	1'200'000
1	98	98^2	850'000
1	85	85^2	970'000

$$X = \begin{bmatrix} 1 & 120 & 120^2 \\ 1 & 132 & 132^2 \\ 1 & 98 & 98^2 \\ 1 & 85 & 98^2 \end{bmatrix}$$

$$y = \begin{bmatrix} 900000 \\ 1200000 \\ 850000 \\ 970000 \end{bmatrix}$$

A higher-order polynomial

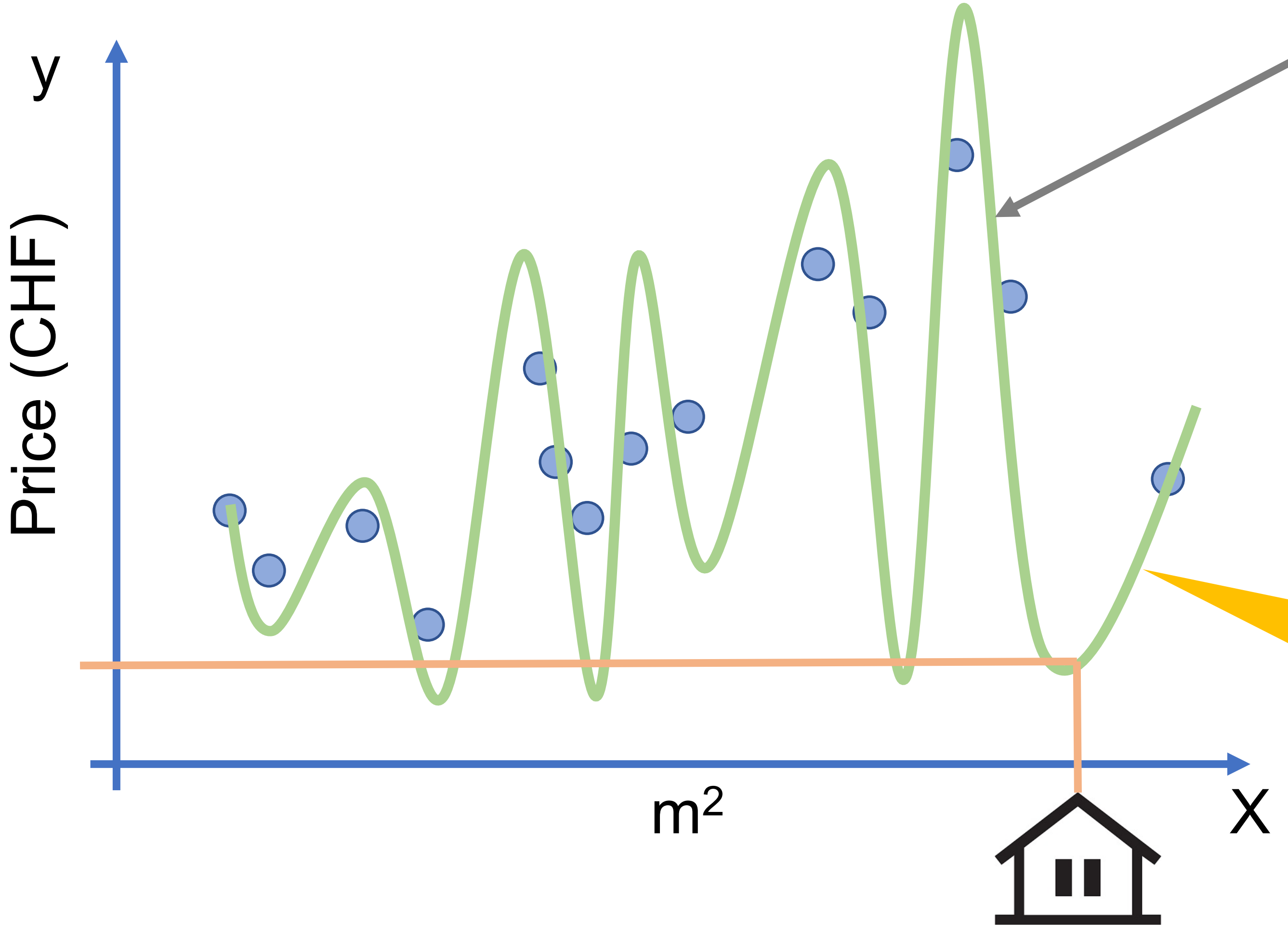


$$f(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_p x^p$$

$\boxed{\text{RSS} = 15}$

But is this really a better model simply because the RSS is smaller?

A higher-order polynomial



$$f(x) = w_0 + w_1 x + w_2 x^2 + \ldots + w_p x^p$$

Is it reasonable to assume that our large house costs so little?

We call this "OVERFITTING". Our model fits well the data we have given it, but does not "generalize" well.

Overfitting

"*Overfitting* is the tendency of data mining procedures to tailor models to the training data, at the expense of generalization to previously unseen data points"

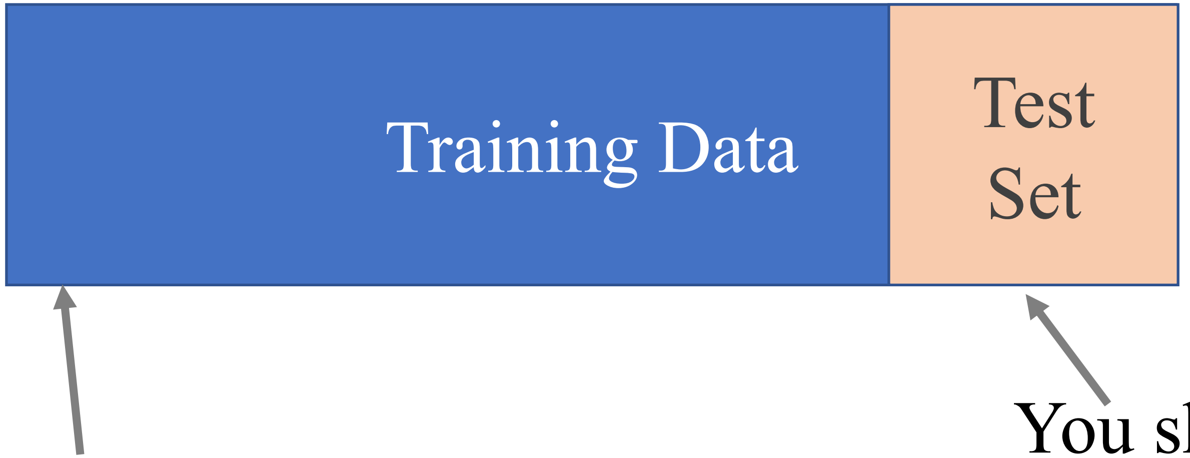
See also Chapter 5 of "Data Science for Business" book.

Evaluation

How do we evaluate how good our model is?

Evaluation of your model

- This is the **most important slide** in this class, and where most people make the **most errors** when modeling data!



Train/Learn model

You should test the prediction error on the test set.

Evaluate error on data you haven't seen!

- We would like to have good predictions, but we don't know what the future brings (nobody does!).

Simulate predictions:
-Remove observations
(create training/test set)
-Fit model on remaining
(use training set)
-Predict on held-out observations
(use test set)

Whole dataset

House area	Lot area	# baths	Sale Price
120	500	…	930k
65	350	…	705k
154	0	…	2010k
220	0	…	3000k
65	15	…	350k
85	35	…	810k
122	0	…	1200k

Evaluate error on data you haven't seen!

- We would like to have good predictions, but we don't know what the future brings (nobody does!).

Simulate predictions:
-Remove observations
(create training/test set)
-Fit model on remaining
(use training set)
-Predict on held-out observations
(use test set)

Training set

House area	Lot area	# baths	Sale Price
120	500	…	930k
65	350	…	705k
154	0	…	2010k
220	0	…	3000k
65	15	…	350k
85	35	…	810k
122	0	…	1200k

Evaluate error on data you haven't seen!

- We would like to have good predictions, but we don't know what the future brings (nobody does!).

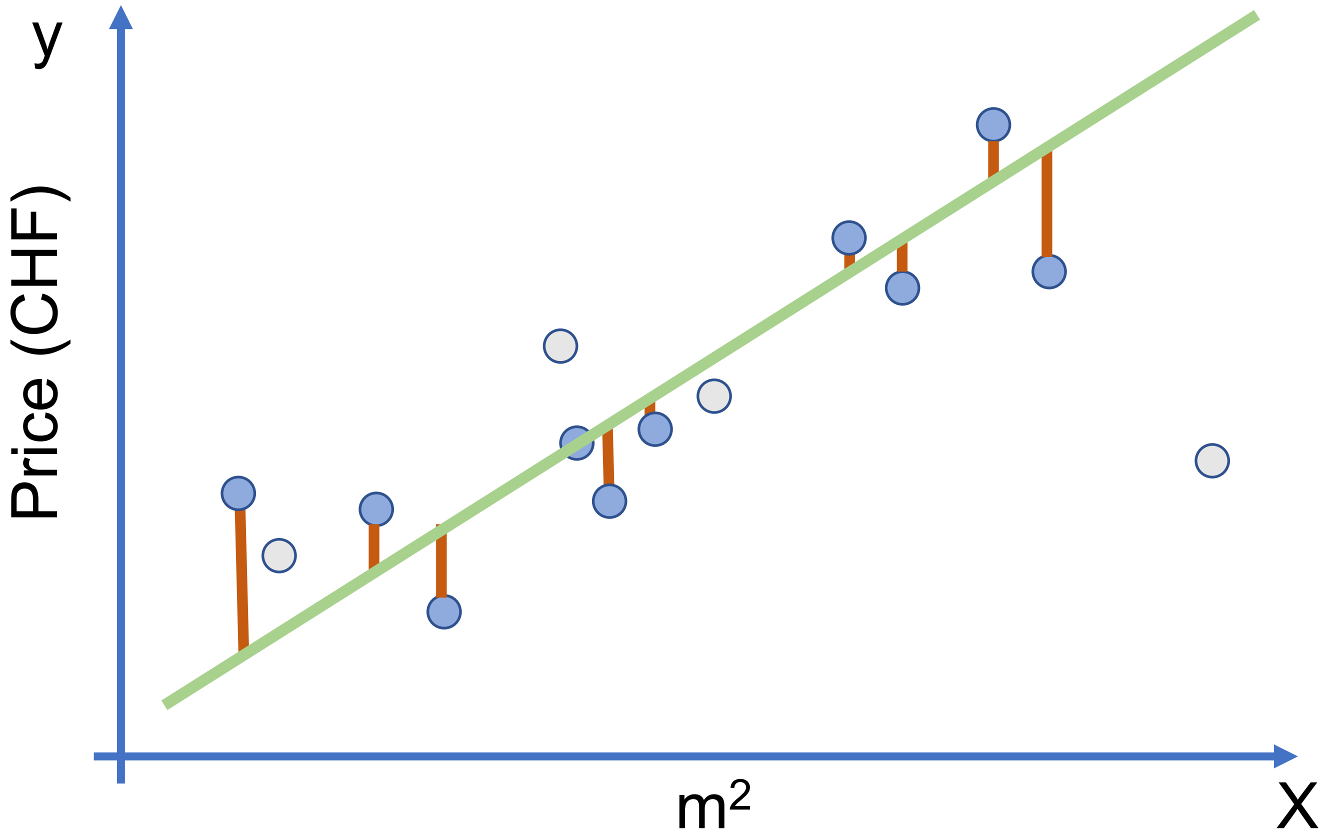
test set

House area	Lot area	# baths	Sale Price
120	500	…	930k
65	350	…	705k
154	0	…	2010k
220	0	…	3000k
65	15	…	350k
85	35	…	810k
122	0	…	1200k

Simulate predictions:
-Remove observations
(create training/test set)
-Fit model on remaining
(use training set)
-Predict on held-out observations
(use test set)

Training Error

Use only training set



Training error(w)=

$(y_{training1}\text{-}f_w(x_{training1}))^2$

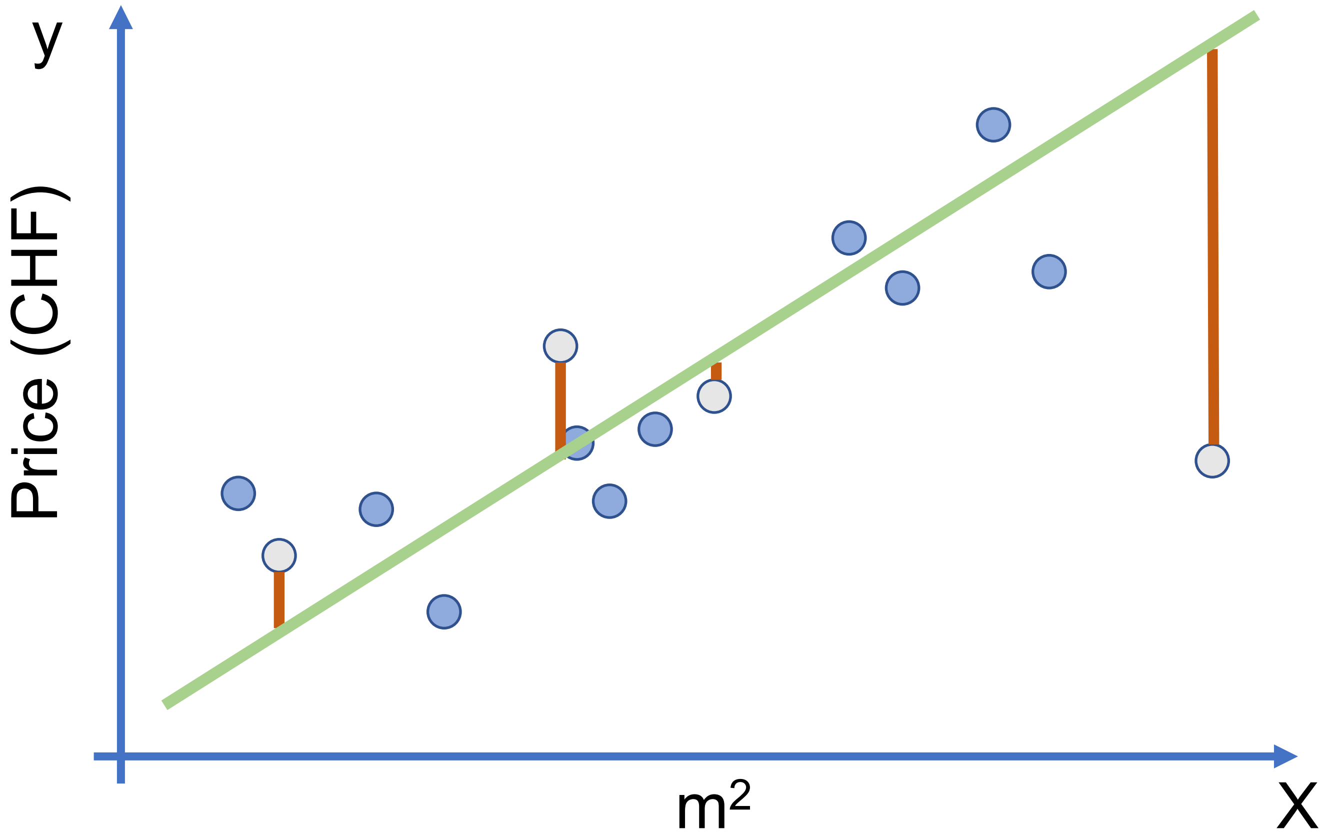
$+ (y_{training2}\text{-}f_w(x_{training2}))^2$

$+ \ldots.$

[for all training examples]

Test Error (a more realistic estimate of the prediction error)

Use only test set



Test error(w)=

$(y_{test1} - f_w(x_{test1}))^2$
$+ (y_{test2} - f_w(x_{test2}))^2$

$+ \ldots.$
[for all test examples]

Assess prediction error using only the test set

80-20 split

- Typically we use an 80-20 split for train/test sets.
- This means we use
  - 80% of the dataset for training the model (learning the parameters w)
  - 20% as the test set for predicting the model accuracy (or error)

A. First we **shuffle** the rows of the dataset.

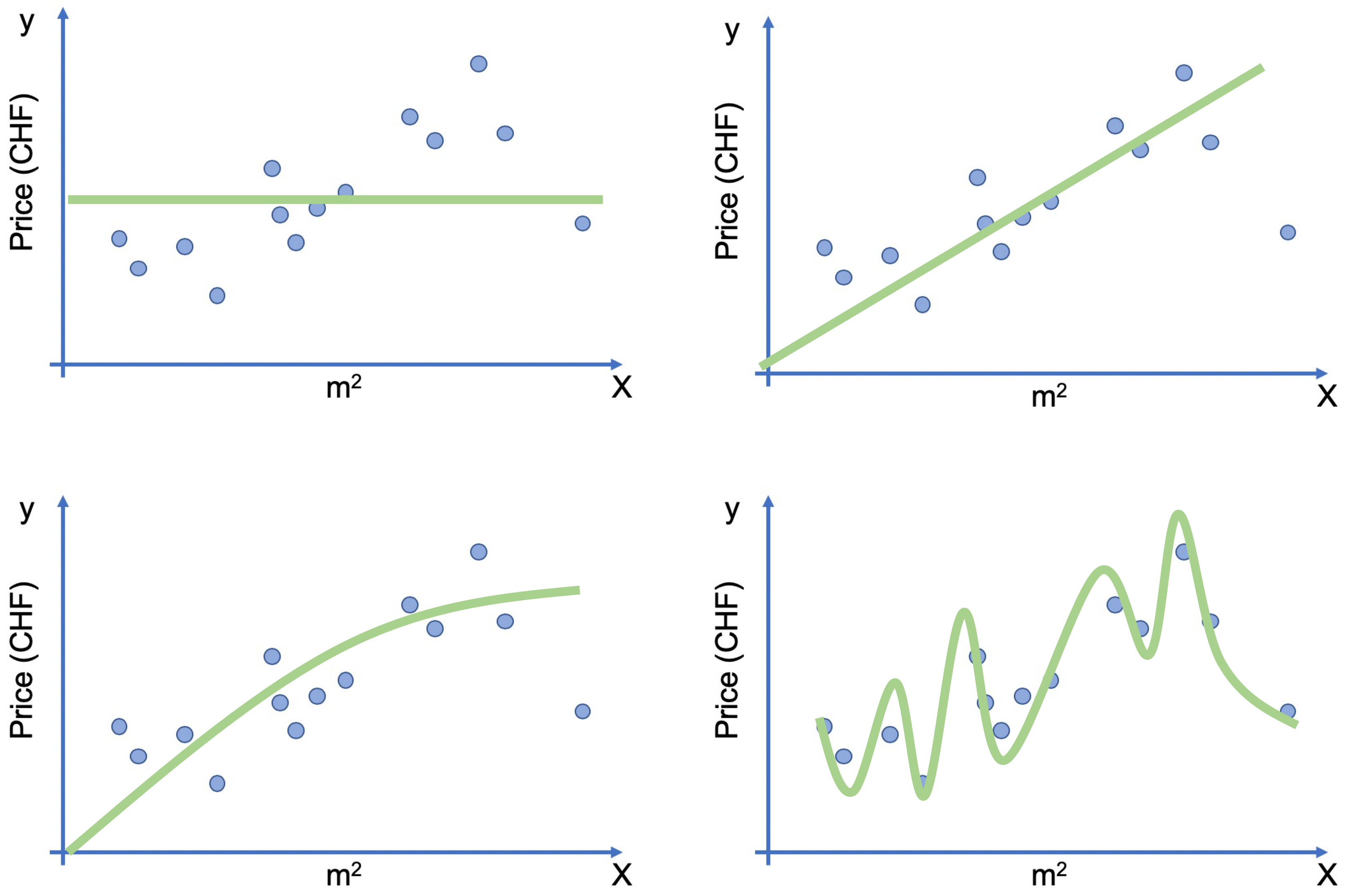
B. We select the first 80% of the rows → train set

C. We select the last 20% of the rows → test set

How to evaluate which model is best?

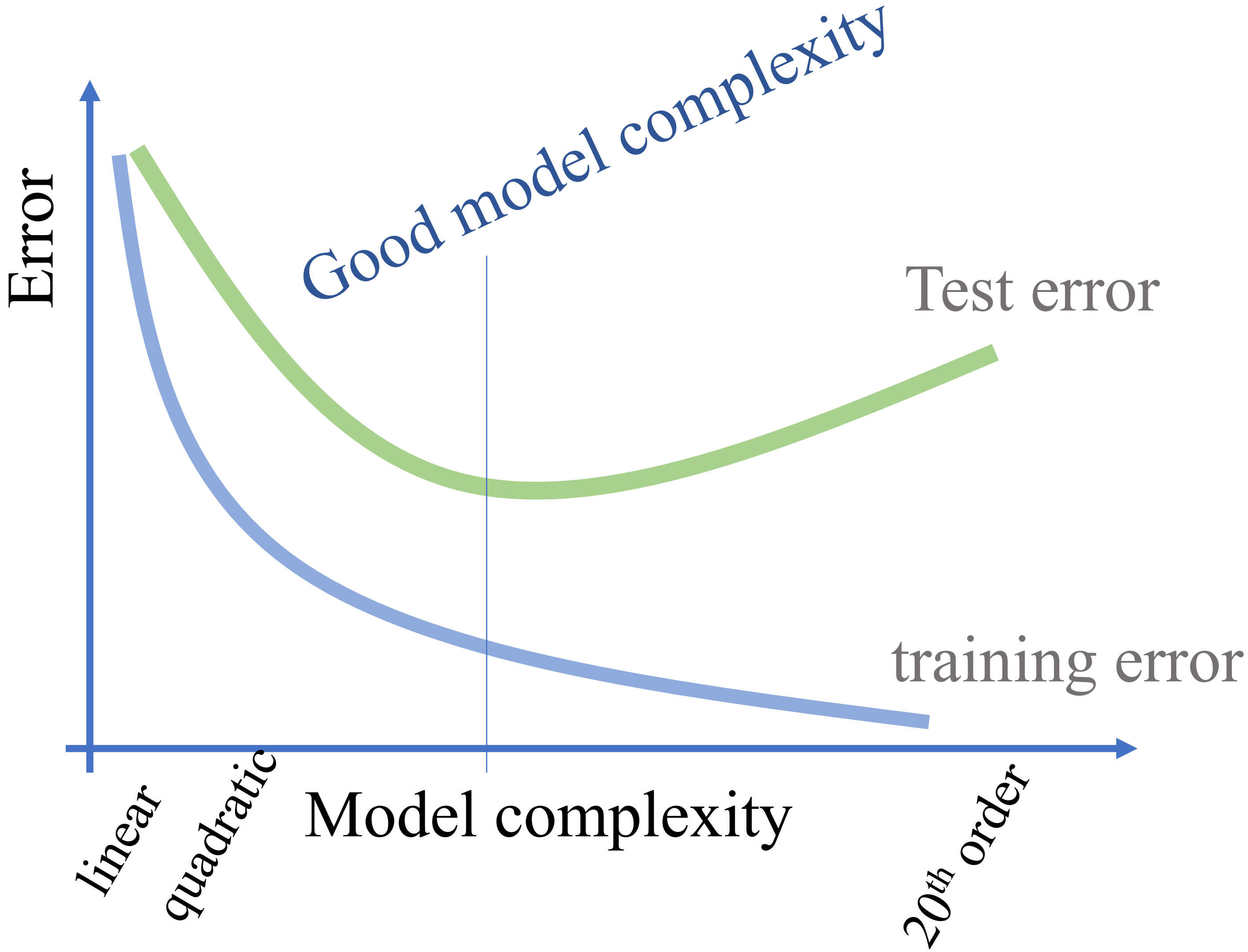
- Imagine we are given many models. How do we find which one we should choose?
- We use again the train/test split and found out which behaves the best for the **test data**.

How to evaluate which model is best?



To select which of those models is the “best”, means which has the lowest error in the **test data,** because this means that it generalizes well.

Train/Test Curves



We pick the model that has the lowest test error. The training error will (almost) always reduce with more complex models. **You want the simplest possible model, with the least test error.**

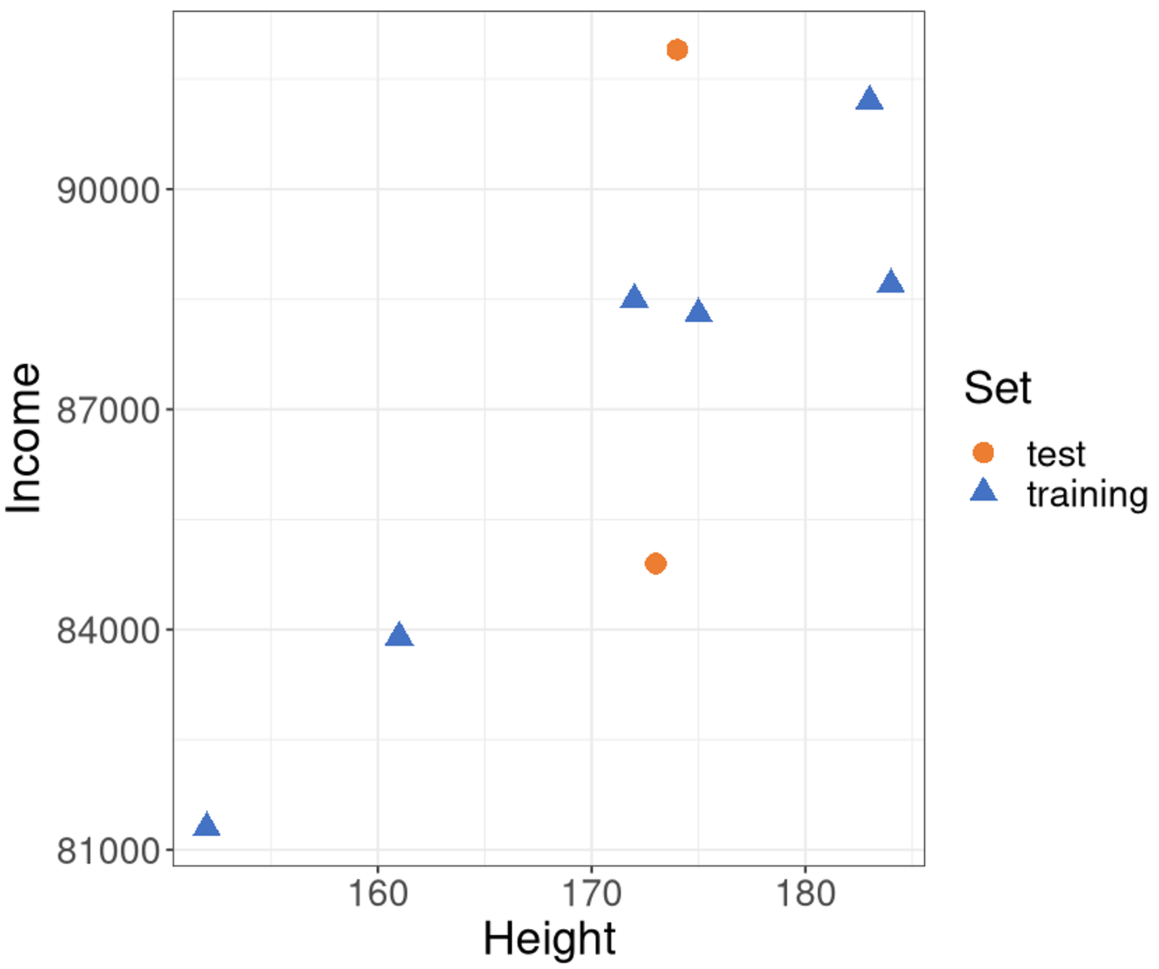
In-class exercise: Income vs Height

We have the following data on income vs height of a person.

1. Compute the MAE of the train and the test data on this model.

$f(x) = 20'000 + 400\,x$

Income	Height
91200	183
83900	161
88500	172
88300	175
81300	152
88700	184
84900	173
91900	174

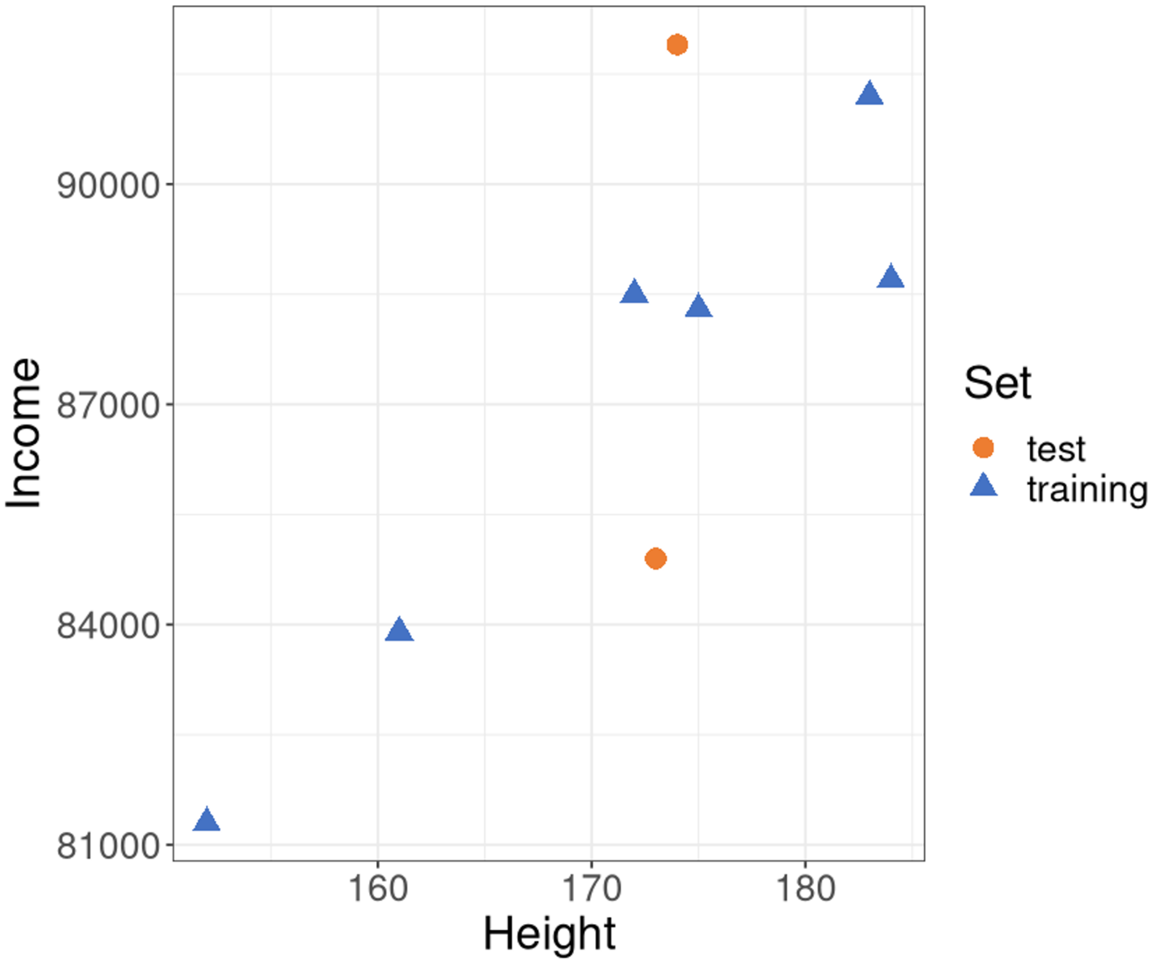


In-class exercise: Income vs Height

We have the following data on income vs height of a person.

1. Compute the MAE of the train and the test data on this model.
$f(x) = 20'000 + 400\,x$
2. Given this new model, $f(x) = 38'600 + 290\,x$, which one would you pick and why?

Income	Height
91200	183
83900	161
88500	172
88300	175
81300	152
88700	184
84900	173
91900	174



Cross-Validation

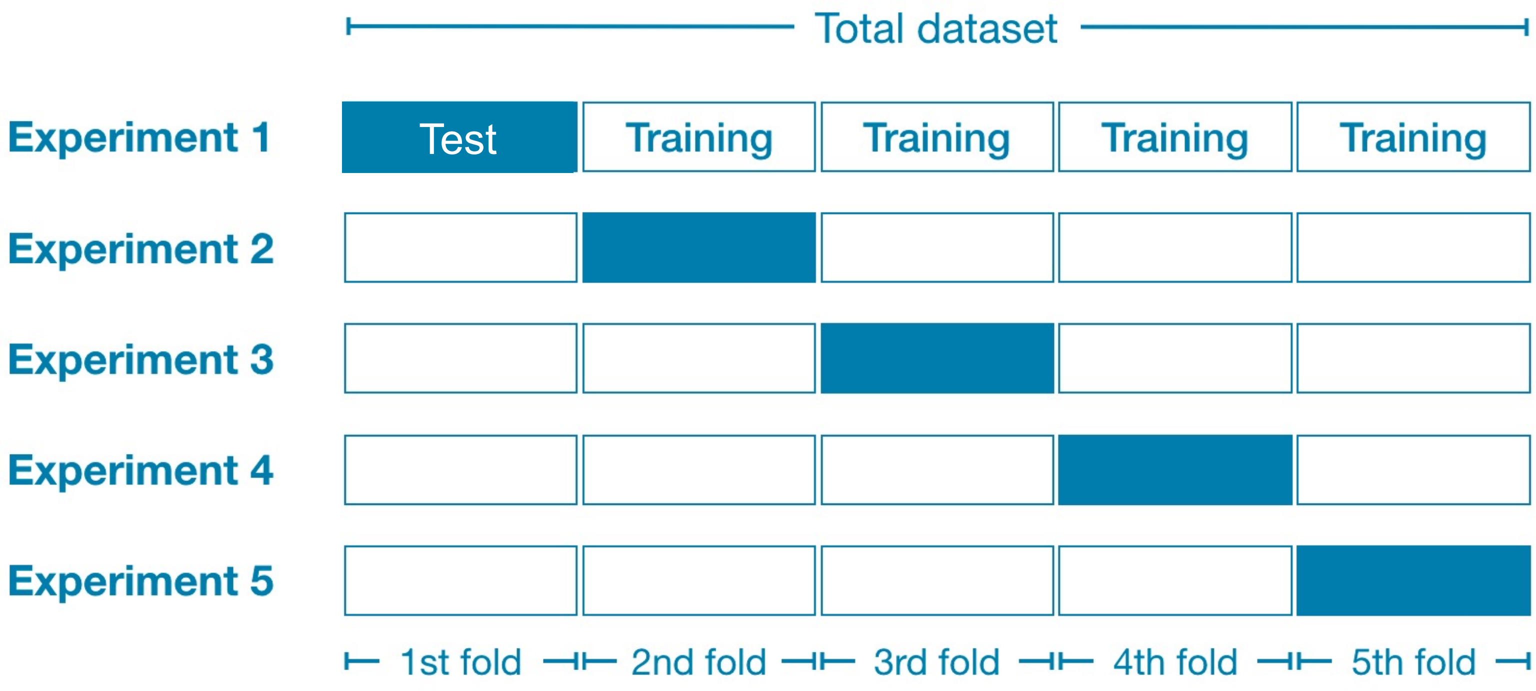
More accurate estimates of the prediction error

Creating many test sets

- If we use only **one test set**, then we may get *good or bad* prediction results depending on what the test set was.
- We can get a more realistic estimate of the prediction error by using **many test sets**, which will yield a better measure of model quality.

K-fold cross validation (CV)

- Typically we use K=5 or 10.
- Test_Error = $\frac{1}{K}\sum_{1}^{K} test\_error_{fold-K}$
- If K=size of dataset, then we have a "leave-one-out" CV. It provides the most accurate estimate of model quality but it is also the most costly.

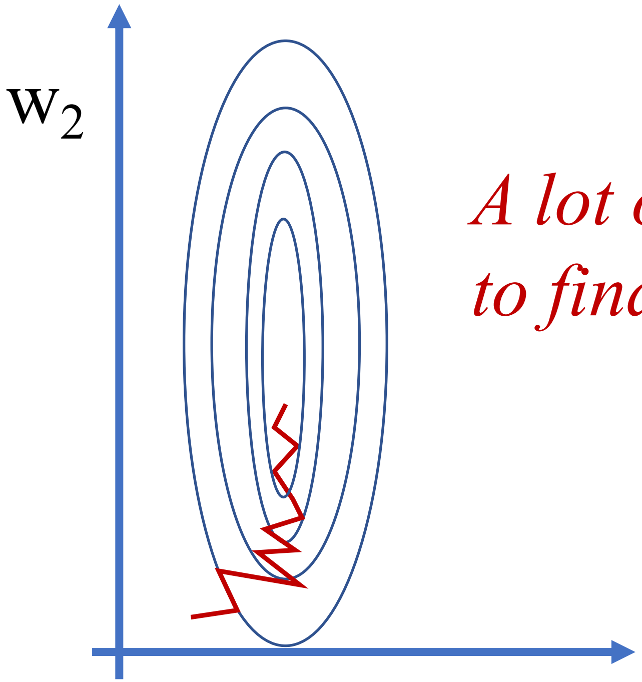


Practical Considerations

Feature Scaling

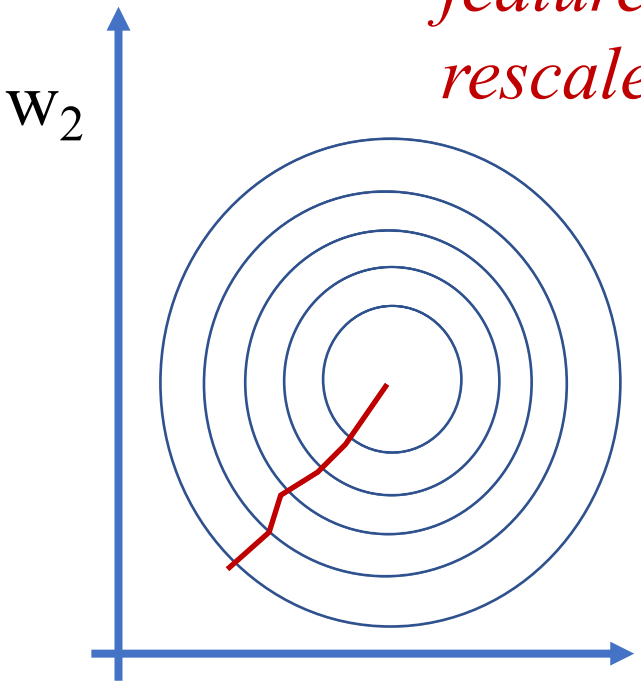
Need to scale features

- Gradient descent may not work as well, when feature values are not in the similar range of values.
- For example,
  - house area = 20-350m^2, but
  - number of bathrooms = 1-5



A lot of oscillations to find the solution

Easier when the features are rescaled



Rescaling also promotes interpretability

- If we don't rescale the features to have the same range (e.g., 0-1) then the weights for the linear regression don't mean much.

Salary = w_0 + **5**years_of_experience + **2**number_of_friends + …

- BUT, we all features have the same range, then the weights give you the **relative importance** of the features.

0-1 scaling

- Goal: to get every feature in the 0-1 range.

- Formula: $x\_new = (x - \min(x)) / (\max(x) - \min(x))$

Mean normalization

- Another possible normalization is to remove the mean value of that feature and divide by some constant (either (max-min), or std).

x_new = (x – mean(x))/(max(x)-min(x))

x_new = (x – mean(x))/std(x)

Now it has zero mean and unit variance = **standardization**

Example: Assume x is the area of a house. Then we have:

x = (size – mean(size)) / (max(size) – min(size))
x = (size – 100)/(350-20) = (size – 100)/330

Practical Considerations

Handing Categorical Variables
1-Hot encoding, Label encoding

Handing Categorical Variables

- So far we assumed the features were numerical.
- Many features will be categorical. How do we deal with those?

Handing Categorical Variables (Binary)

Categorical Binary (2 values) → convert into 0,1

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	PaymentMethod	Monthl
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	Electronic check	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Mailed check	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Mailed check	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Bank transfer (automatic)	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	Electronic check	
5	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	Electronic check	
6	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	Credit card (automatic)	

Handing Categorical Variables (>2 values)

Electronic check, Mailed Check,
Bank transfer, Credit card,…

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	PaymentMethod	Monthl
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	Electronic check	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Mailed check	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Mailed check	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Bank transfer (automatic)	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	Electronic check	
5	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	Electronic check	
6	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	Credit card (automatic)	

Handing Categorical Variables (>2 values)

1-hot encoding

Electronic check, Mailed Check,
Bank transfer, Credit card,…

PAY_ECHECK	PAY_MCHECK	PAY_BANK	PAY_CREDIT
1	0	0	0
0	1	0	0
0	1	0	0
0	0	1	0
1	0	0	0
1	0	0	0
0	0	0	1

PaymentMethod	Monthl
Electronic check	
Mailed check	
Mailed check	
Bank transfer (automatic)	
Electronic check	
Electronic check	
Credit card (automatic)	

Handing Categorical Variables (>2 values)
Label encoding

Electronic check, Mailed Check,
Bank transfer, Credit card,…

Electronic check: 1
Mailed check: 2
Bank transfer: 3
Credit card: 4

…

PAYMENT_METHOD	PaymentMethod	Monthl
1	Electronic check	
2	Mailed check	
2	Mailed check	
3	Bank transfer (automatic)	
1	Electronic check	
1	Electronic check	
4	Credit card (automatic)	

1-Hot encoding vs Label encoding

Rule of thumb:

- When there are 10-15 distinct categorical values use 1-hot-encoding
- When there are more distinct categorical values use label encoding.

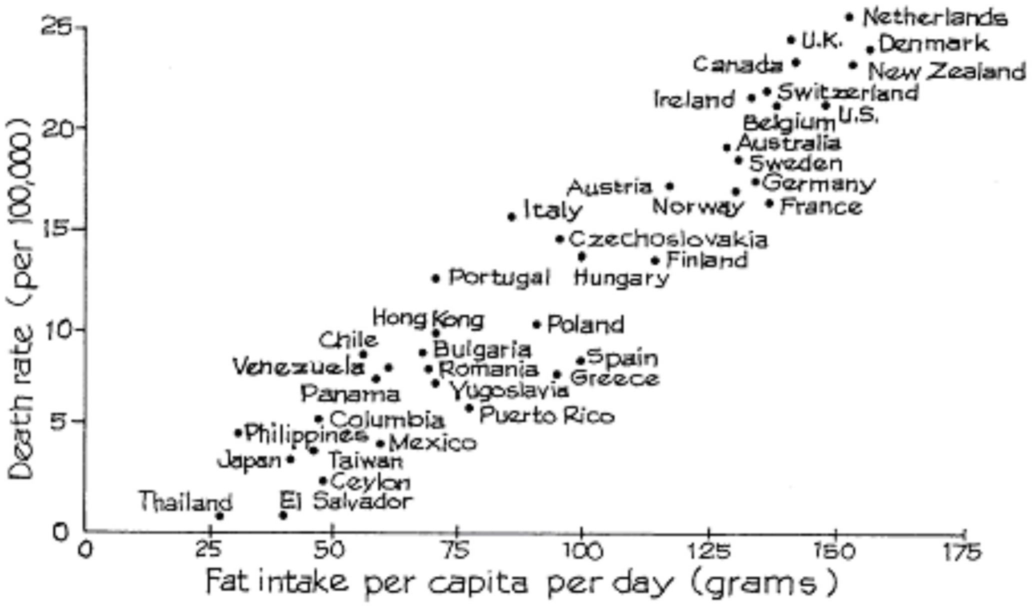
- In practice, you try both and see which one gives lower **test error. You use that one!**
- See also here an example:
  - https://www.kaggle.com/alexisbcook/categorical-variables

Correlation and Causation

Correlation/Association vs Causation

- When we find that some variables are associated, we should not fall into the trap of misinterpreting it as **causation.**

Figure 8. Cancer rates plotted against fat in the diet, for a sample of countries.



Source: K. Carroll, "Experimental evidence of dietary factors and hormone-dependent cancers," *Cancer Research* vol. 35 (1975) p. 3379. Copyright by *Cancer Research*. Reproduced by permission.

Does fat intake cause cancer? This graph seems to provide such evidence.

Correlation/Association vs Causation

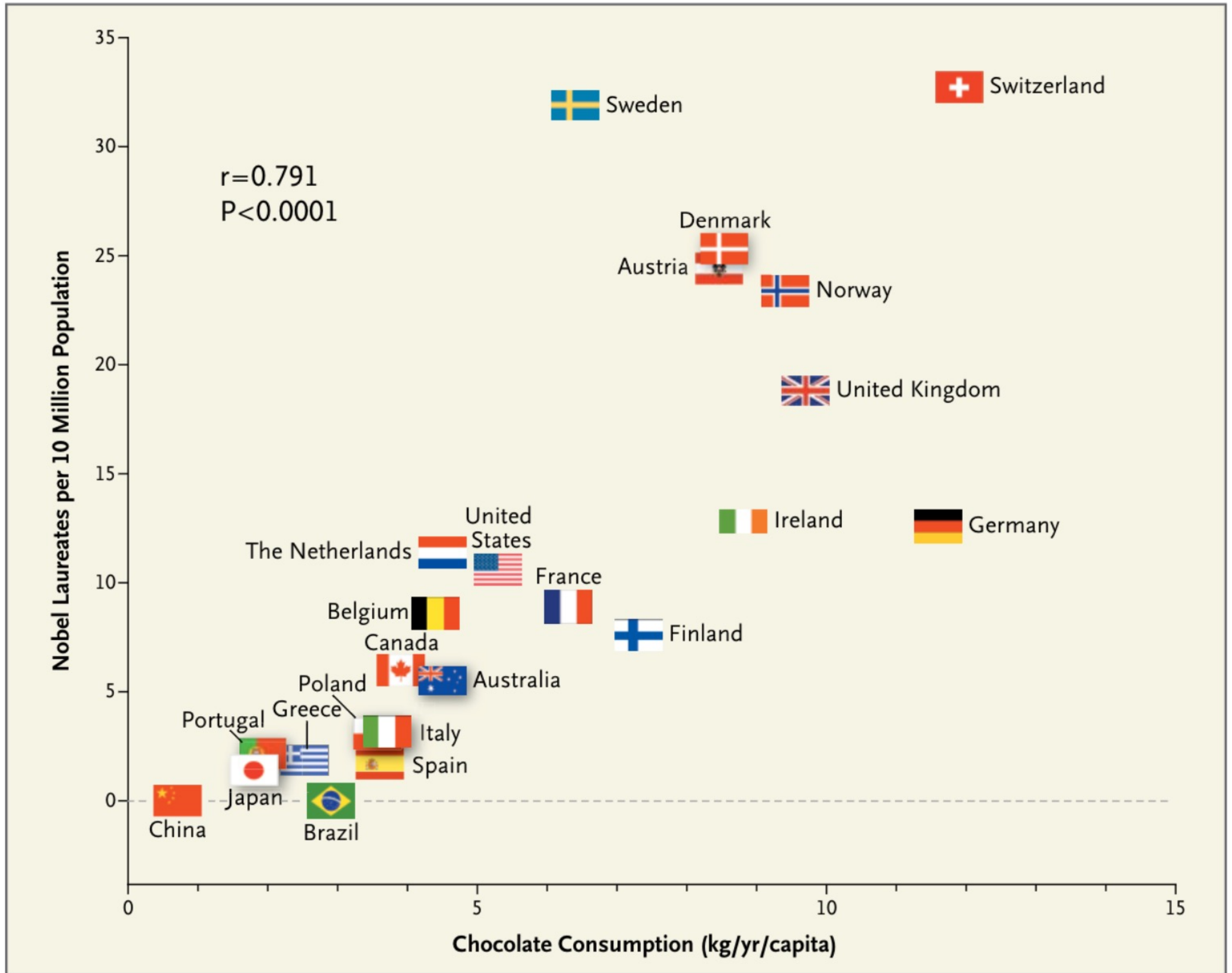
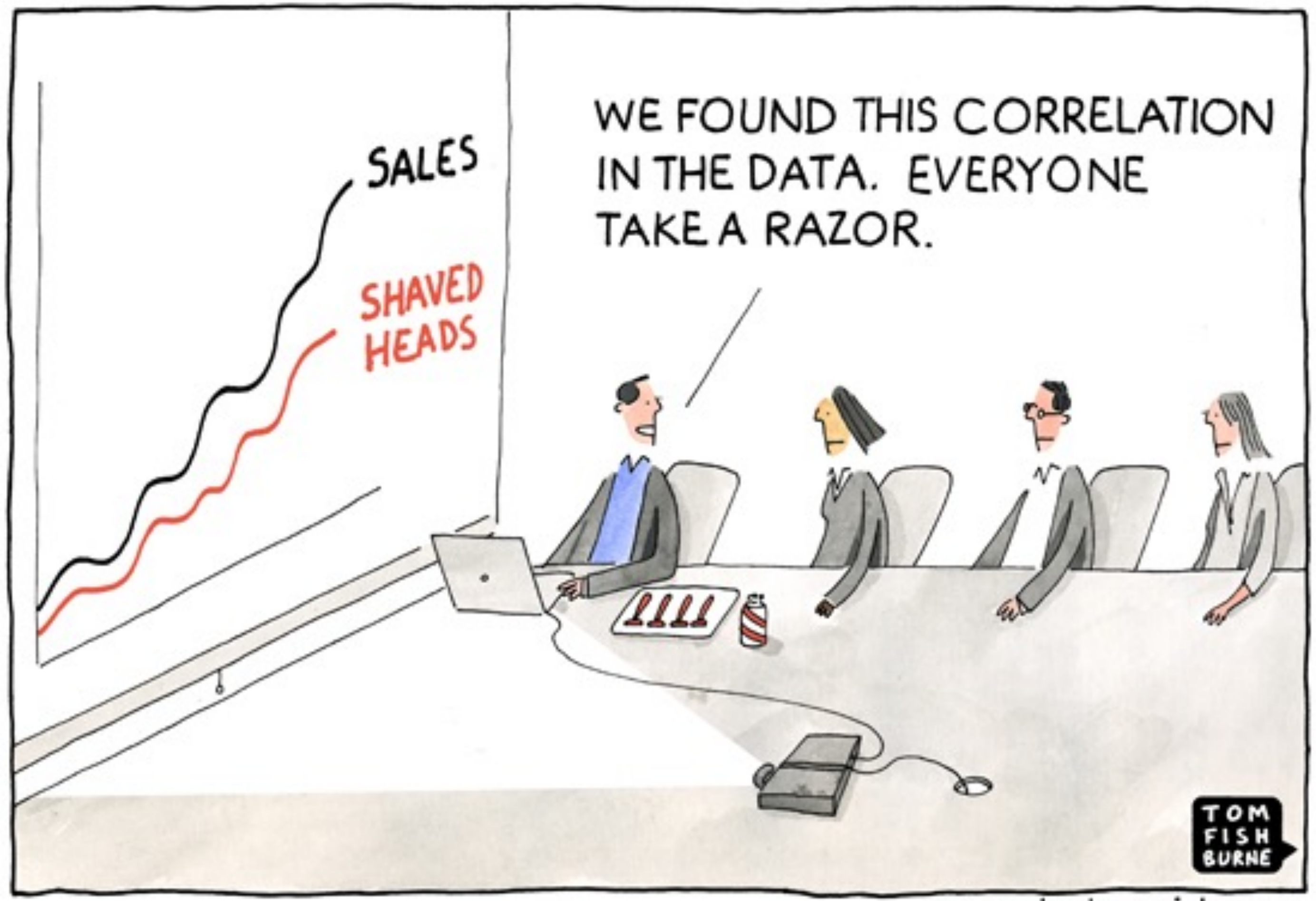
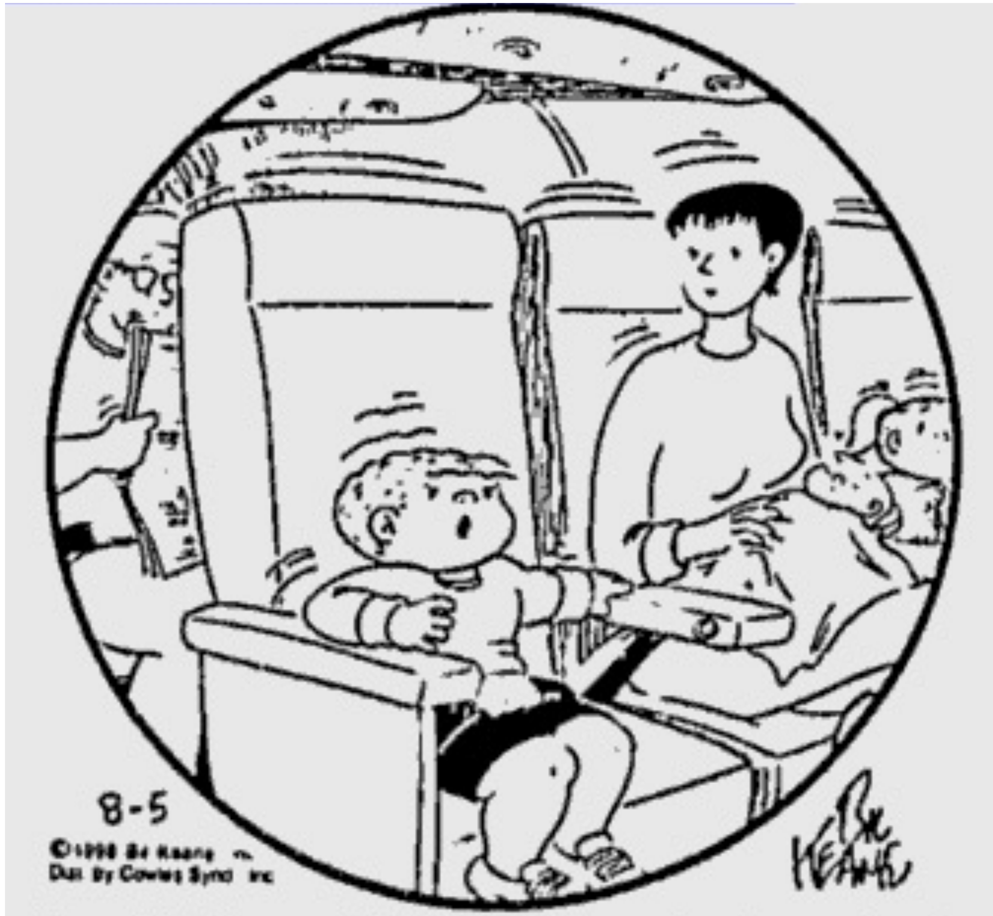


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

High-correlation between chocolate consumption and Nobel Laureates in a country…Switzerland rocks!



© marketoonist.com



"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

Correlation and Causation

Regression suggests that there is a correlation (especially for large R^2 values)

- **Correlation** – Values track each other
  - Height and Shoe Size
  - Grades and SAT Scores
- **Causation** – One value directly influences another
  - Education Level → Starting Salary
  - Smoking → Cancer

Correlation and Causation

Correlation does not imply causation

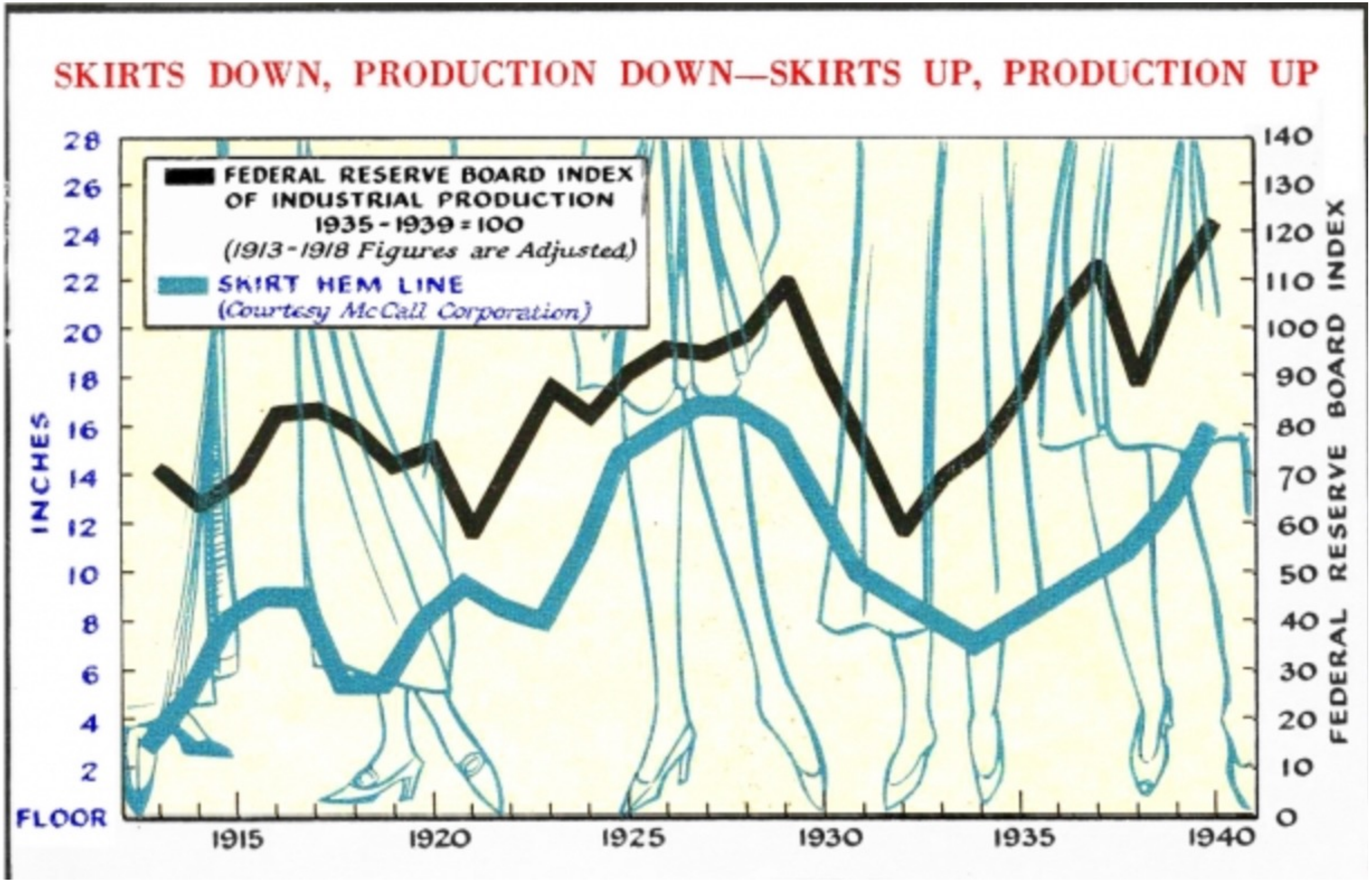
- Correlation can be result of causation from a hidden “confounding variable”
- A and B are correlated because there’s a hidden C such that C→A and C→B

❖Homeless population and crime rate
Confounding variable: unemployment
❖Forgetfulness and poor eyesight
Confounding variable: age

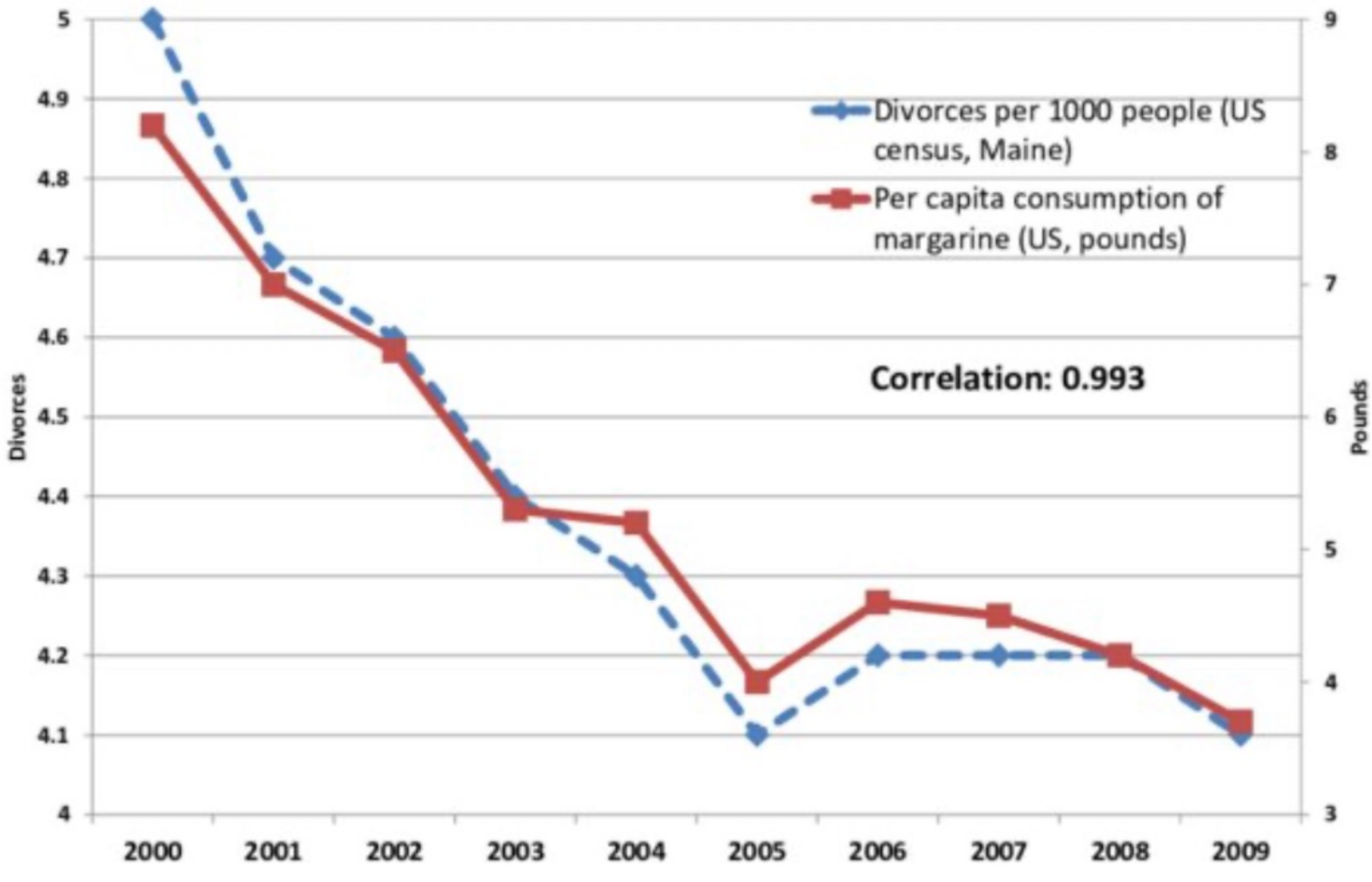
Group Activity

- Find some more cartoons that highlight the distinction between correlation and causation!

Surprising correlation #1



Surprising correlation #2



So remember. Correlation does not imply causation!

Terminology

Terminology

- Regression
- Linear regression
- Multi-linear regression
- Train(ing) error, test error
- Train/Test curves
- K-fold cross validation
- Rescaling
- 1-hot encoding
- Label encoding
- Causation vs correlation